Supplementary Materials for

Khoe-San genomes reveal unique variation and confirm deepest population divergence in *Homo sapiens*

Carina M. Schlebusch[#*,1,2,3], Per Sjödin*[1], Gwenna Breton*[1], Torsten Günther[1], Thijessen Naidoo[1,2,4], Nina Hollfelder[1,5], Agnes Sjöstrand[1], Jingzi Xu[1], Lucie Gattepaille[1], Mário Vicente[1], Douglas Scofield[3,6], Helena Malmström[1,2], Michael de Jongh[7], Marlize Lombard[2], Himla Soodyall[8,9], Mattias Jakobsson[#1,2,3]

Address:
Human Evolution, Department of Organismal Biology, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden

* Equal contribution
# Correspondence: carina.schlebusch@ebc.uu.se; mattias.jakobsson@ebc.uu.se

Table of Contents

1. Sampling, DNA handling and sequencing

A total of 25 Khoe-San samples from five different populations (Karretjie People, Nama, Ju|'hoansi, |Gui and ||Gana, and !Xun) were sequenced. DNA samples from individuals were collected with the subjects' informed consent. The project was reviewed and approved by the University of Witwatersrand (South Africa) Human Research Ethics Committee (M180654), the Swedish Ethical Review Authority (Dnr 2019-05174) and the South African San Council. A description of sample groups, group membership, number of individuals, place of sampling and origin are outlined in Table S1.1. For a full description and historical background of groups see Supplementary Materials for (Schlebusch et al. 2012).

DNA from EDTA-blood was extracted using the salting-out method (Miller et al. 1988) . DNA was quantified using a nanodrop, and diluted to 50ng/µl with ddH2O.

1.1. WGA libraries preparation and sequencing

Whole Genome Amplification (WGA) was performed with the IllustraGenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Uppsala, Sweden) according to kit instructions, but with modifications to the protocol as outlined in (Schlebusch et al. 2012).

Duplicate sequencing libraries for each sample were prepared at SciLife Lab SNP & Seq platform in Uppsala, Sweden. Libraries were prepared from 1 µg DNA using the TruSeq DNA sample prep kit V2 (cat# FC-121-2001/2002, Illumina Inc.) targeting an insert size of ~380 bp. The library preparation was performed according to the manufacturers' instructions (guide#15026486 rev A). The quality of the libraries was evaluated using the Bioanalyzer from Agilent Technologies (DNA 1000 Chip). The adapter-ligated fragments were quantified by qPCR using the Library quantification kit for Illumina (KAPA Biosystems) on a StepOnePlus instrument (Applied Biosystems/Life technologies) prior to cluster generation and sequencing. Each library was sequenced on two lanes on Illumina HiSeq 2000 machines (100 bp paired-end reads) at the SciLife Lab SNP & Seq platform in Uppsala.

1.2. Genomic libraries preparation and sequencing

DNA libraries were prepared from 1 µg of DNA extract using the TrueSeq® DNA Sample preparation kit v2 (cat#FC-121-2001/2002, Illumina Inc.) targeting an insert size of 380bp. The libraries were constructed according to the manufacturer´s instructions (guide#15026486 rev C). Each library was sequenced on one Illumina HiSeq2000 lane (100 bp paired-end reads) at the SciLife Lab SNP & Seq platform in Uppsala, except for the library of sample KSP063, which was sequenced on two lanes.

Sampling, DNA handling and sequencing Tables and Figures

Table S1.1: A description of sample groups, including group names, group membership, place of sampling and origin, and number of individuals. Country abbreviations: AN – Angola, BT – Botswana, NM – Namibia, SA – South Africa. Adapted from Table S1 in (Schlebusch et al. 2012).

Ethnic group name	Main population group	Place of sampling (country)	Place of origin (if different from sampling)	Number of genomes sequenced
Karretjie People	San	Colesberg (SA)		5
Nama	Khoe	Windhoek (NM)		5
|Gui and ||Gana	San	Kutse Game Reserve (BT)		5
Ju|'hoansi	San	Tsumkwe (NM)		5
!Xun	San	Omega camp (NM), Schmidtsdrift (SA)	Around Menongue (AN)	5
Total				25

2. Read Mapping and Bam processing, SNP and indel calling and filtering

The processing steps described in sub-sections 2.1 to 2.5 are summarized in Figure S2.1.

2.1. Mapping

BAM files were generated by mapping the reads to the 1000 genomes phase 2 reference assembly (human reference genome GRCh37/hg37, hs37d5.fa.gz, dated 2011-07-07, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/) with bwa 0.6.2 mem (Li and Durbin 2009) using default parameters. The resulting BAM files were merged to one file per sample with Picard (Broad Institute) and all reads were marked with read group information containing flowcell id, library id and lane.

2.2. BAM processing

BAM files were further processed in GATK v.2.5.2 (DePristo et al. 2011) and Picard v.1.92. The post-processing of BAMs involved: duplicate marking in Picard, realignment around indels in GATK, calculating the MD flag with samtools (Li et al. 2009) *calmd* and Base Quality Score Recalibration (BQSR) in GATK. The BQSR was done in two rounds for each sample; the first round of BQSR used one file as "knownsites" to mask known variants (dbsnp_137.b37.vcf downloaded from ftp://ftp.broadinstitute.org//bundle/2.5/b37/) and four covariates (ReadGroupCovariate, CycleCovariate, ContextCovariate, QualityScoreCovariate). After this first BQSR round SNPs and indels were called on both the processed BQSRed BAM files and the processed non-BQSRed BAM files resulting in four VCFs files (dbsnp-recalibrated-snp, dbsnp-recalibrated-indel, non-recalibrated-snp, non-recalibrated-indel). These resultant four VCFs were used as additional "knownsites" input files for the second round of BQSR. The second round BQSR thus used five VCF files as "knownsites" (dbsnp_137.b37.vcf, and four called VCFs) and the same four covariates as the first round. The input BAMs for the second round of BQSR were the non-BQSRed processed BAMs. The two rounds of BQSR were done because dbSNP variants might not sufficiently represent Khoe-San variation and thus using only dbSNP variants as a mask in the BQSR step might cause decreased base qualities for novel Khoe-San specific variants.

2.3. Initial SNP calling to generate VCF files to use as "knownsites" in BQSR (see above)

Initial SNP calling was done using the UnifiedGenotyper module in GATK v.2.5.2. Group-calling of all 25 sample BAMs were done together (per chromosome). SNPs and indels were called separately. The dbsnp_137.b37.vcf file was used to annotate SNPs with dbSNP names during the SNP calling. The resultant VCF files were used in the second round of BSQR (see above).

2.4. Final SNP calling and processing of "KSP only dataset"

BAM files used in the final SNP call only passed through the second round of BQSR described above. Final SNP calling was done using UnifiedGenotyper module in GATK v.2.5.2. Three different SNP calls were done on the input BAMs. Using the "read_group_black_list" flag in GATK, SNP calls were done separately for 1) Whole Genome Amplified (WGA) libraries, 2) Genomic libraries and 3) All libraries combined. For all three SNP calls, group-calling was done on 25 samples together, dbsnp_137.b37.vcf file was used to annotate SNPs with dbSNP names during the SNP calling, SNPs were called separately from indels, a strand call confidence of 20.0 was used, all sites present in reference genome were emitted (not just variant sites), and VCFs were extensively annotated (SpanningDeletions, Coverage, DepthPerAlleleBySample, QualByDepth, FisherStrand, MappingQualityRankSumTest, ReadPosRankSumTest, GCContent, HaplotypeScore, HomopolymerRun, TandemRepeatAnnotator, VariantType).

Following the three separate SNP calls the VCFs were subjected to three separate Variant Quality Score Recalibrations (VQSRs). VQSR was done in GATK v.2.8.1 and annotations considered during VQSR were: Coverage, QualByDepth, FisherStrand, MappingQualityRankSumTest and

ReadPosRankSumTest. To train the machine learning algorithm we used the following known and truth sites with the parameters: dsa (known=false, training=true, truth=true, prior=15.0), 1000G (known=false, training=true, truth=false, prior=5.0), dbsnp (known=true, training=false, truth=false, prior=2.0). The “dsa” dataset was obtained from Illumina Omni 2.5M SNP typing on the same 25 samples, as described in (Schlebusch et al. 2012). Only sites variable in the 25 samples were selected and 70% of the dataset were selected randomly to be used in VQSR. The rest (30%) were kept to do quality checks. The “1000G” dataset corresponds to sites identified in the 1000 genomes project and was downloaded from ftp://ftp.broadinstitute.org//bundle/2.5/b37/. A “tranche” level of 99.9 during VQSR was used to mark SNPs for filtering.

After the separate SNP calling and VQSRs on the WGA, Genomic and All libraries VCFs, information from the WGA and Genomic VCF (and to a lesser extent the All libraries VCF) were combined into a consensus SNP VCF using the program samla (https://github.com/douglasgscofield/samla) in which the gwa-ksp METHOD (--method gwa-ksp) was employed. A full description of which options are included in this method and the selection criteria that these options entail are described on the samla github page.

Indels called on “All libraries combined” were recalibrated using GATK v.3.1.1. Annotations considered for VQSR were: FisherStrand, Depth, QualByDepth, MappingQualityRankSumTest and ReadPosRankSumTest. The algorithm was trained using the dataset recommended in GATK: Mills_and_1000G_gold_standard.indels.b37.vcf (known=true,training=true,truth=true,prior=10.0). A “tranche” level of 99.9 was used in the ApplyRecalibration step to mark indels for filtering.

2.5. Quality control

We obtained the coverage of the final BAMs with QualiMap/2.2 (Okonechnikov et al. 2015) using one of two options:

1-including reads marked as duplicates: qualimap bamqc -bam input.bam --java-mem-size=40G -nt 5 -nr 500 -nw 300 -sd -sdmode 0 -c -outdir outfolder

2-not including reads marked as duplicates: qualimap bamqc -bam input.bam --java-mem-size=40G -nt 5 -nr 500 -nw 300 -sdmode 0 -c -outdir outfolder

We compared the genotypes obtained after processing of the sequencing data described in sub-section 2.4 to the SNP array data available for the 25 samples (Illumina HumanOmni2.5 (Schlebusch et al. 2012)). We used only the 30% of the variants which were not used in the “dsa dataset” at the BQSR step, which corresponds to ~455,000 SNPs. After renaming the SNPs to “chromosome:position”, we used plink/1.90b4.9 (Purcell et al. 2007) “--bmerge --merge-mode 6” and flipped variants with mismatching alleles once to account for strand error.

2.6. Combined SNP calling and processing of “KSP and HGDP dataset”

The processing steps described in this sub-section are summarized in Figure S2.2.

BAM files of 11 HGDP samples (see section 3.1) were SNP group-called together with the 25 KSP BAM files (processing described in section 2.1 to 2.3) to form a combined KSP+HGDP group SNP called VCF file. SNP calling was done using the UnifiedGenotyper module in GATK v.2.5.2. SNPs and indels were called separately and the dbsnp_137.b37.vcf file was used to annotate SNPs with dbSNP names during the SNP calling. A strand call confidence of 30.0 was used, all sites present in the reference genome were emitted (not just variant sites) and VCFs were extensively annotated (SpanningDeletions, Coverage, DepthPerAlleleBySample, QualByDepth, FisherStrand, MappingQualityRankSumTest, ReadPosRankSumTest, GCContent, HaplotypeScore, HomopolymerRun, TandemRepeatAnnotator, VariantType).

VQSR was done in GATK v.2.8.1 and annotations considered during VQSR were: Coverage, QualByDepth, FisherStrand, MappingQualityRankSumTest and ReadPosRankSumTest. To train the machine learning algorithm we used the following known and truth sites with the parameters: dsa (known=false, training=true, truth=true, prior=15.0), 1000G (known=false, training=true,

truth=false, prior=5.0), dbsnp (known=true, training=false, truth=false, prior=2.0). The “dsa” dataset was obtained from Illumina Omni 2.5M SNP typing on the same 25 samples, as described in Schlebusch et al 2012. Only sites variant in the 25 samples were selected and 70% of the dataset were selected randomly to be used in VQSR. The rest (30%) were kept to do quality checks. The “1000G” dataset corresponds to sites identified in the 1000 genomes project and was downloaded from ftp://ftp.broadinstitute.org//bundle/2.5/b37/. A “tranche” level of 99.9 during VQSR was used to mark SNPs for filtering.

Furthermore the KSP+HGDP dataset was also filtered for SNPs that failed a VQSR procedure on a SNP call on each separate dataset. This step was done to eliminate any dataset specific problematic SNPs that remain in the data. The SNP call and VQSR on the separate datasets followed the same parameters as on the combined dataset. Hardy-Weinberg Equilibrium (HWE) filtering was performed by setting to missing SNPs that contain only heterozygous individuals (with four missing individuals allowed) in the separate HGDP and KSP datasets. This combined KSP+HGDP group-called, quality filtered VCF, was merged with further comparative data (see section 3).

Read mapping and BAM processing, SNP and Indel calling and filtering Tables and Figures

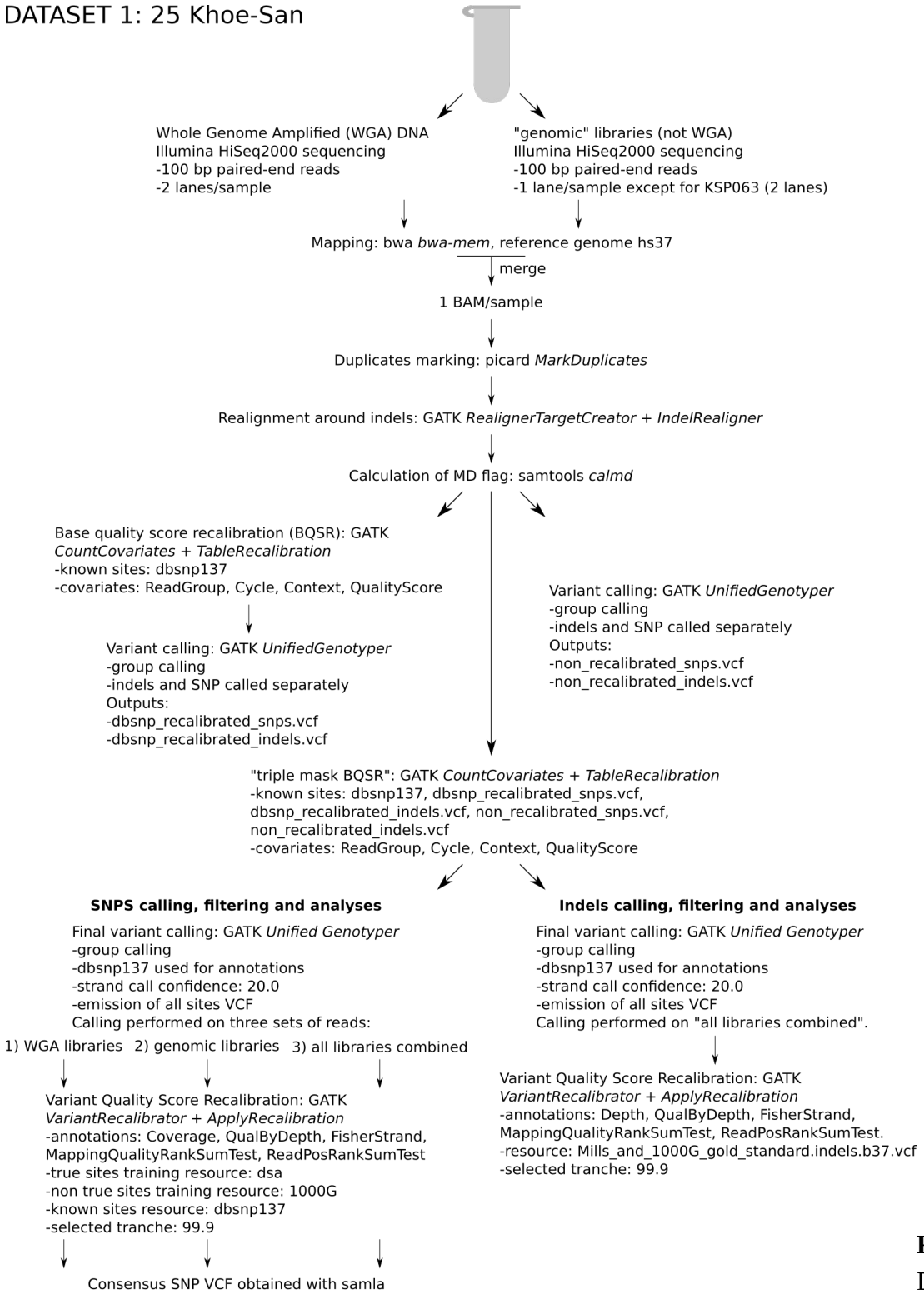
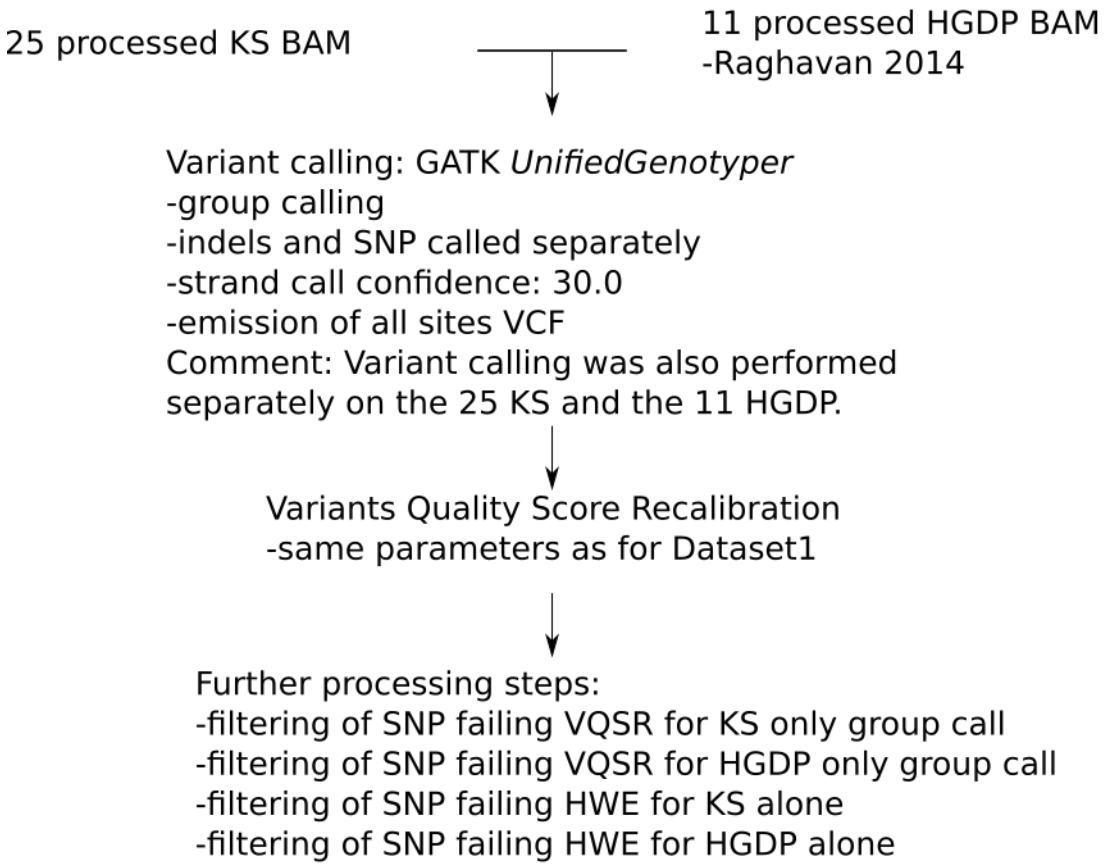


Figure S2.1: Description of the generation of the "Khoe-San dataset", "KSP dataset", or DATASET 1.



Further modifications include filtering for HWE, filtering for missingness, phasing etc.
Analyses with this dataset include: count of variants, heterozygosity estimates, Venn diagram of private and shared variants, allele sharing between pairs of populations, masking of non Khoe-San ancestry, PSMC, MSMC, TT method, GPhoCS, PBS scan, etc.

Figure S2.2: Description of the generation of the "KS+HGDP dataset", “KSP+HGDP dataset”, or DATASET 4.

3. Comparative data processing and merging

3.1. HGDP data (HGDP)

BAM files of 11 HGDP samples (one each of Dai, Dinka, French, Han, Karitiana, Mandenka, Mbuti, Papuan, San, Sardinian, and Yoruba) aligned to the human reference genome build 37.1 were obtained (http://www.cbs.dtu.dk/suppl/malta/data/Published_genomes/bams/). The data was originally generated by (Meyer et al. 2012) and the re-mapping and generation of the BAM files performed and described in (Raghavan et al. 2014). The 11 HGDP BAM files were used and SNP group called together with the 25 KSP BAM files (processing described in sub-sections 2.1 to 2.3). The KSP+HGDP group-calling and VQSR is described in sub-section 2.6.

3.2. Complete Genomics diversity set (CG)

22 samples (five ASW: African-American SW; four GIH: Gujarati; four MKK: Maasai in Kinyawa, Kenya; five MXL: Mexican; four TSI: Toscani) from the Complete Genomics diversity set (Drmanac et al. 2010) were downloaded as VCF format files from ftp://ftp2.completegenomics.com/vcf_files/Build37_2.0.0/ (version number of the Complete Genomics assembly software: CGAPipeline_2.0.0.26). The files were then processed to keep only SNP sites, half-called genotypes were set as missing and missing stretches were extended to one line per missing position. Only individual genotype information was extracted and sites marked as VQLow were set as missing. Sites below a depth of 10 and a genotype quality of 30 were set to missing. HWE filtering was performed by setting to missing SNPs that contain only heterozygous individuals in the dataset (with four missing individuals allowed). Individual VCFs were merged into one VCF using vcf-merge (Danecek et al. 2011). All sites were kept to allow for merging with other comparative data.

3.3. 1000 Genomes data typed on Complete Genomics platform (KGP)

30 samples (five CEU: Northern and Western European ancestry in Utah; five CHS: Southern Han Chineses; five LWK: Luhya; five PEL: Peruvians; five PJL: Punjabi; five YRI: Yoruba) from the 1000 Genome project (Auton et al. 2015), sequenced to high coverage and typed on the Complete Genomics platform were downloaded as VCF format files from http://www.1000genomes.org/announcements/complete-genomics-data-release-2013-07-26. The files were then processed to keep only SNP sites, half-called genotypes were set as missing and missing stretches were extended to one line per missing position. Only individual genotype information was extracted and sites marked as VQLow was set as missing. Sites below a depth of 10 and a genotype quality of 30 were set to missing. HWE filtering was performed by setting to missing SNPs that contain only heterozygous individuals in the dataset (with four missing individuals allowed). Individual VCFs were merged into one VCF using vcftools vcf-merge (Danecek et al. 2011). All sites were kept to allow for merging with other comparative data.

3.4 Data from (Lachance et al. 2012) (LC)

Data from 15 samples (five Hadza; five Sandawe; five western rainforest hunter-gatherers: three Baka, one Ba.Kola and one Bedzan) were obtained directly from Joseph Lachance as Complete Genomics Mastervar files. The software cgatools (http://cgatools.sourceforge.net/) was used to filter (varType=del,varType=ins,varType=sub) and convert Mastervar to VCF format (using the --include-no-calls flag). The files were then processed to keep only SNP sites, half-called genotypes were set as missing and missing stretches were extended to one line per missing position. Only individual genotype information was extracted and sites marked as VQLow was set as missing. Sites below a depth of 10 and a genotype quality of 30 were set to missing. HWE filtering was performed by setting to missing SNPs that contain only heterozygous individuals in the dataset (with four missing individuals allowed). Individual VCFs were merged into one VCF using vcftools vcf-merge (Danecek et al. 2011). All sites were kept to allow for merging with other comparative data.

We sometimes use the terms "Pygmy" or “Pygmies” to refer to the individuals from three populations from western rainforest hunter-gatherers (Baka, Ba.Kola and Bedzan) and one population of eastern rainforest hunter-gatherers (Mbuti). However we recognize the derogatory connotation of these terms and refer when possible to the names of the individual populations.

3.5 Generation of final merged comparative dataset - “global dataset”

The processing steps described in this sub-section are summarized in Figure S3.1.

The CG, KGP and LC data was merged with the KSP+HGDP group-call dataset using vcftools vcf-merge. All sites were kept to allow for further merging. Thereafter variant VCF files were generated and only variant unfiltered sites were kept. The dataset was furthermore filtered for SNP missingness to allow no more than 10% missing data.

3.6 Archaic genomes

Additionally the Neandertal and Denisova genomes were prepared for comparative data analysis. The Denisova genome (published originally in (Meyer et al. 2012) and remapped in (Raghavan et al. 2014)) was obtained for this study from http://www.cbs.dtu.dk/suppl/malta/data/Published_genomes/bams/. The Neandertal genome (Prüfer et al. 2014) was obtained from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/bam/. SNP calling on the BAMs was done using the UnifiedGenotyper module in GATK v.2.5.2. SNPs were called per chromosome using default settings and the dbsnp_137.b37.vcf file was used to annotate SNPs with dbSNP names during the SNP calling. All sites present in reference genome were emitted and VCFs were extensively annotated (SpanningDeletions, Coverage, DepthPerAlleleBySample, QualByDepth, FisherStrand, MappingQualityRankSumTest, ReadPosRankSumTest, GCContent, HaplotypeScore, HomopolymerRun, TandemRepeatAnnotator, VariantType).

3.7. Ancient San genome: Ballito Bay A (BBA)

We included the data for a high-coverage ancient San individual (dated to ~2,000 years ago), the Ballito Bay A boy. Details of the processing and diploid genotype calling are found in (Schlebusch et al. 2017). To allow merging with our comparative dataset, sites with DP<10, GQ<30 and/or a LowQual flag were set to missing.

3.8. Six additional Khoe-San genomes from Simon’s Genome Diversity Project (SGDP)

The all-sites VCFs for six additional Khoe-San genomes - two ‡Khomani and four Ju|'hoansi - from the Simon’s Genome Diversity Project (Mallick et al. 2016) were obtained from https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public/. To allow merging with our comparative dataset, sites with DP<10, GQ<30 and/or a LowQual flag were set to missing.

3.9. Generation of the “extended global dataset”

The Ballito Bay A and SGDP samples VCFs were merged with the combined dataset described in sub-section 3.5, using vcf-merge (vcf-tools version 0.1.13 (Danecek et al. 2011)) by chromosome (only autosomes). All sites were kept.

A relatedness analysis was performed with KING (Manichaikul et al. 2010). Preparation of the input data involved 1-merging the individual chromosome VCF with GATK/3.5.0 CatVariants, 2-obtaining a plink TPED file with vcftools --plink-tped option, 3-converting the tped into a plink binary fileset with plink/1.90b4.9. Two pairs of first-degree relatives (estimated “Kinship” coefficient > 0.177) were identified: a pair of Ju|'hoansi samples, one from our own dataset (KSP116) and one from the SGDP dataset (S_Ju_hoan_North-2); and a pair of Maasai samples from the CG dataset (NA21732 and NA21737). The Ju|'hoansi pair was expected because the samples in our dataset and in the SGDP dataset come from the same sample collection, collected by Prof. Trefor Jenkins (University of the Witwatersrand) in Tsumkwe, Namibia. For both pairs of samples

the sample with the highest missingness (respectively S_Ju_hoan_North-2 and NA21737) was excluded from the all-sites VCF using GATK/3.5.0 SelectVariants and the -xl_sn option. The final combined dataset contains 108 samples. Finally, the VCF was filtered for missingness. The filter for sites genotyped in less than 90% of the samples (and not already filtered for another reason) was set to FAIL5.

Comparative data processing and merging Tables and Figures

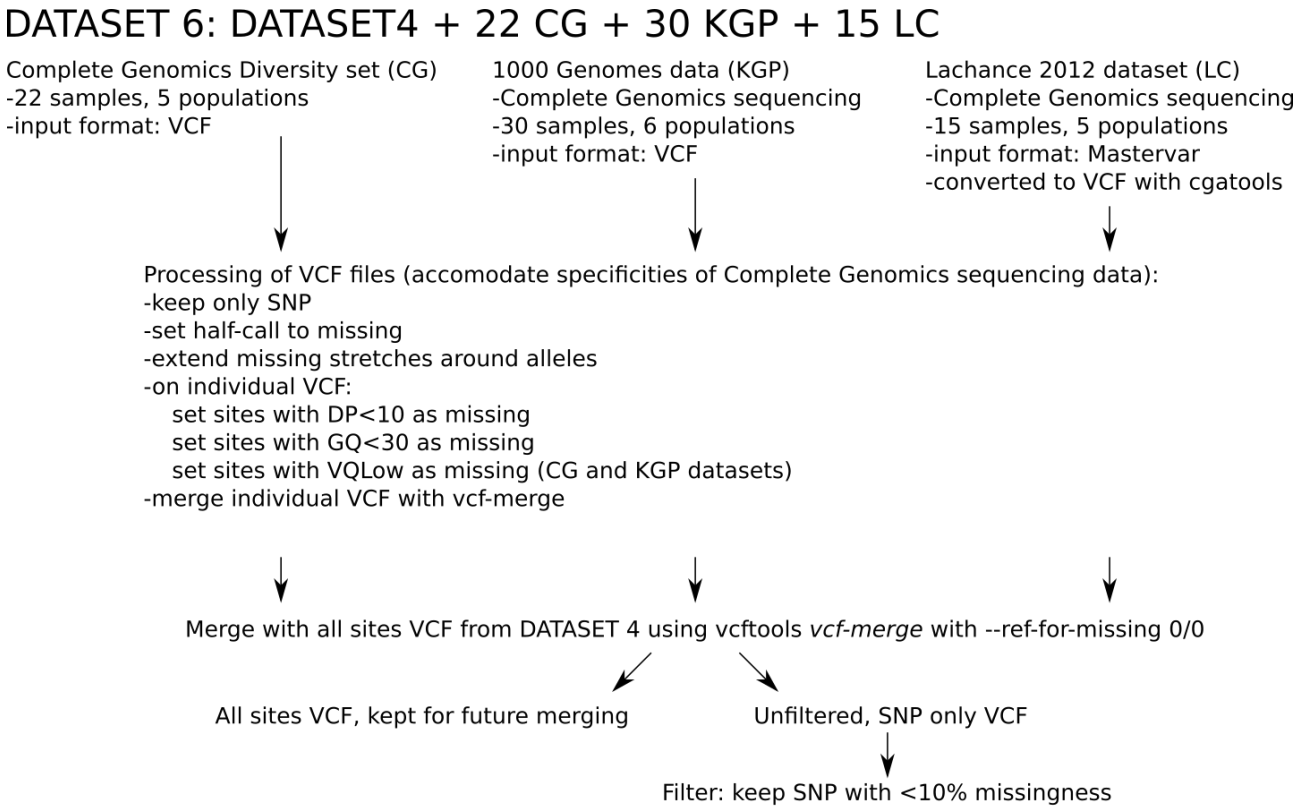


Figure S3.1: Description of the generation of the "global dataset" or DATASET 6.

4. Preparation for data analysis

4.1. Phasing and imputation

Two phasing strategies were followed. Statistical phasing and imputation of VCF files was done using Beagle (v. 4.0) with 25 iterations (Browning and Browning 2007). Statistical phasing and imputation of plink files was performed using fastPhase (v. 1.4.0) (Scheet and Stephens 2006). Assumed missingness was set to 10%. Chromosomes were phased separately and at 25 iterations each (flags: -T25, number of random starts of EM algorithm; -C25, number of iterations of EM algorithm, -H100, number of haplotypes sampled; -K25, number of clusters).

4.2. Ancestral state inference

The human ancestral variant were determined using three outgroups: Chimpanzee (panTro4, UCSC, downloaded from ftp://hgdownload.soe.ucsc.edu/goldenPath/panTro4/), Gorilla (gorGor3, UCSC, ftp://hgdownload.soe.ucsc.edu/goldenPath/gorGor3/), and Orangutan (ponAbe2, UCSC, ftp://hgdownload.soe.ucsc.edu/goldenPath/ponAbe2/). We only used sites for which the ancestral state was confidently called (the three great apes showed the same variant and no missingness, and at most one additional variant among the individuals in the analyzed dataset).

4.3. Preparation of recombination maps

The YRI recombination map (HapMap release 24) and CEU+YRI combined recombination map (HapMap release 24) were downloaded from http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/. The genetic map was smoothed according to the procedure described in (Schlebusch et al. 2015).

4.4. Masking

We used the RFMix software (Maples et al. 2013) to identify specific ancestries of genomic fragments for the Khoe-San individuals. The data was imputed with SHAPEIT (Delaneau et al. 2012). We ran the RFMix analyses with three parental groups; San, Other African and Eurasian. We selected the five Ju|'hoansi, the HGDP San and the five Karretjie individuals as sources for the San ancestry. Six YRI, one Mandenka, five 5 LWK, four MKK and one Dinka individual were selected to represent the “Other African” ancestry. Finally, five CEU, one French, one Sardinian, four TSI, four GIH and five PJL were labelled as Eurasian source. We ran RFMix analyses with two extra iterations to account for admixture in the source populations and to minimize assignment errors, we set three minimum reference haplotypes per tree node and a window size of 0.02 cM. We used the YRI genetic map as recombination map. Positions in the genomes outside the map windows were excluded. Source populations were kept and we kept only regions which both alleles were assigned to be San specific.

4.5. Additional filtering for specific analysis

Additional filtering for repeat regions, poorly called regions, known Copy Number Variants (CNVs) and CNVs called in our datasets was performed for certain analyses and are described in their specific sections below.

5. Variant summaries

5.1. Autosomal genome summaries

Various genome wide counts and statistics regarding the 22 autosomes of the 25 Khoe-San individuals sequenced in this study are displayed in Table S5.1. Main Table 1 contains a subset of these statistics.

After mapping and quality procedures (sections 3 and 4), we called a total of 20,020,719 SNPs in the autosomes and 19,001,637 remained after VQSR procedures. After further group-wide quality filtering on the called SNPs, 18,637,959 autosomal dinucleotide SNPs remained of which 1,960,665 (10.5% of the dataset) were novel variants (compared to dbSNP built 151). Around 29% of the variants were singletons. 3.2% of the total dataset were non-singleton novel variation. The two southern Khoe-San groups, Nama and Karretjie People, had the largest fraction of novel variants (Table S5.1). Individuals had an average of 5,079,953 SNPs per individual (standard deviation: 46,214) of which around 0.65% were exonic variants.

The average sequencing depth with duplicates removed was 53.4x (range: 45.1 - 59.8 per individual). Individual values for the coverage and the concordance rate with the SNP array data are given in Table S5.2. Two samples (one Karretjie and one Nama) had high variations in coverage along the genome (standard deviation of respectively 166.7x and 277.5x). Thus, these two individuals were excluded from some of the analyses. The uncalled or missing positions compared to the reference genome was below 1.4% in all individuals. The average individual concordance rate was 98.87% (range: 96.81 - 99.19%). The lowest concordance rates were observed in the two samples with the highest variations in coverage. The average concordance rate when they are excluded was 99.04%.

5.2. Heterozygosity

Heterozygosity of each individual was determined by direct counting of heterozygous sites divided by the number of sites in the autosomal reference genome (after subtracting missing and filtered sites). The heterozygosity estimates were calculated for the “Khoe-San only” dataset (see sub-section 2.4 for preparation) and for the group called dataset of Khoe-San together with 11 HGDP samples (see sub-section 2.6 for preparation). Both estimates are plotted in Figure S5.1 and the mean heterozygosities of the Khoe-San dataset are displayed in Table S5.1 and Main Table 1 (“Mean Heterozygosity (Genomic)”).

In general the estimates were higher for the Khoe-San only dataset (ranging from 12.4E-4 to 13.1E-4). The !Xun had the highest mean heterozygosity (12.9E-4) and the Ju|'hoansi the lowest (12.6E-4). In the group-call dataset the Khoe-San heterozygosities (11.3E-4 to 11.9E-4) were lower than in the Khoe-San only dataset, however, all Khoe-San individuals still had higher heterozygosity estimates compared to any comparative individual. The HGDP San individual’s heterozygosity (11.3E-4) was in the lower range but compared well with the Ju|'hoansi in our dataset (11.4E-4 to 11.5E-4). The HGDP San individual is from the Ju|'hoansi population and our Ju|'hoansi samples were collected in the same collection that the HGDP San were collected – see (Schlebusch et al. 2012) Supplementary Materials. The highest non-Khoe-San value was 10.9E-4 for the Mandenka. In general non-African heterozygosities (ranging from 6.0E-4 in the Karitiana to 8.3E-4 in the French) were much lower than those of Africans (10.6E-4 in the Dinka to 11.9E-4 in the !Xun).

The estimates from the group-called dataset of Khoe-San together with 11 HGDP samples are plotted in Main Figure 1C; the mean values for Khoe-San (excluding the HGDP San individual), other Africans, and non Africans are shown. The standard error of the mean per individual, averaged across individuals, was calculated in each group. The values are: 0.057% for the Khoe-San, 0.06% for the other Africans, and 0.071% for the non Africans.

5.3. Runs of Homozygosity (RoH)

Runs of Homozygosity were created using plink (v. 1.07) with the following parameters: sliding windows of 50 SNPs, allowing one heterozygous site per window, with an overlapping proportion of 0.05, final window sizes of at least 200 kb and 200 SNPs with a minimum SNP density of one in 20 kb and a gap of 50 kb between SNPs before the run of homozygosity is split in two.

Khoe-San groups had the lowest mean total RoH lengths for all RoH length classes (Figure S5.2A). Although other Africans had on average higher mean total RoH, there are differences among the diverse groups (Figure S5.2C) with rain-forest hunter-gatherers displaying much higher RoH than western and eastern Africans in the shortest mean total RoH class. When mean total RoH for the shortest class is sorted descendingly per individual (Figure S5.2B), Khoe-San individuals in general have the shortest mean total RoH, while Mandenka and Dinka also have relatively short mean total RoH, Yoruba and Mbuti have longer mean total RoH and non-Africans the longest. Among Khoe-San groups, !Xun and Karretjie have the lowest mean total RoH in the shortest length class.

Main Figure 1C shows the total ROH length –for ROH up to 1 Mbp – averaged across individuals in the Khoe-San (excluding the HGDP San individual), other Africans, and non Africans. The standard error of the mean of individuals is also displayed. It is 7.031% for the Khoe-San, 19.521% for the other Africans and 22.905% for the non Africans. The high standard errors in the “other Africans” and in the “non Africans” could be explained by recent inbreeding and / or the diversity of ancestral backgrounds in these groupings.

5.4. Site frequency spectra

The counts of known and novel variants in every frequency class was determined by vcftools (Danecek et al. 2011) (counts function) and plotted as a histogram in R (Team 2013) (Figure S5.3A). The novel variants are according to dbSNP v.151. In general the lower allele classes (singletons, doubletons, etc) had a larger percentage of novel variants than the more common variants. Only 0.03% of the variants in the allele class of 25 were novel variants. The unfolded site frequency spectrum is shown in Figure S5.3B.

5.5. Allelic diversity, Shared variation and Private allelic diversity

Shared variation between different groupings of individuals was determined using the vcf-compare module of vcftools and visualized using the R package “VennDiagram” (Figure S5.4). For comparisons between Khoe-San, Other Africans and Non-Africans, sample sizes were correlated by selecting the following samples from the “Khoe-San and HGDP” group-call dataset: four Khoe-San individuals (one individual from each of the Karretjie, Nama, !Xun and Ju|'hoansi groups), four Other Africans (one Mbuti, one Dinka, one Mandenka, one Yoruba) and four Non-Africans (one Papuan, one Sardinian, one Han, one Karitiana). For the comparisons between the different Khoe-San groups the Khoe-San only dataset was used and here no adjustments were made as all groups contain the same number of individuals (five per group).

The proportion of private SNPs in the Khoe-San was higher than in Other Africans (35% vs 25.3% of all SNPs), while Khoe-San and Other Africans shared 39.7% of SNPs. When Non-Africans are included only 24.9% of variants are shared between all three groups. It appears that Khoe-San share more variants with Non-Africans (5.77%) compared to Other Africans sharing with Non-Africans (4.68%), which could be due to recent Non-African admixture in the southern Khoe-San groups (see sub-section 6.7). The Khoe-San and Other Africans together (eight individuals) share 10.6% variation that is not present in the four Non-African individuals.

For the five Khoe-San groups, 27.4% of variant sites are shared between all five groups, while variants private to each group ranged from 5.86% (Ju|'hoansi) to 6.98% (!Xun). For pairwise shared variation, in general more variants were shared between the two southern Khoe-San groups (Nama and Karretjie share 1.97% of the total number of variant sites) and the two northern Khoe-San groups (!Xun and Ju|'hoansi share 2.28% of total variant sites) (Figure S5.4 third row). The central | Gui and ||Gana group shared intermediate amounts of variation with all the other groups (1.54-

1.7%) and very similar amounts with two northern Khoe-San groups (sharing 1.52% of variation with northern Khoe-San) compared to southern Khoe-San groups (sharing 1.55% variation with southern Khoe-San). Southern groups had more private novel variation (Figure S5.4 bottom row) than northern and central groups. This could possibly be explained due to the facts that: 1-the Ju|'hoansi were previously included in genome studies (HGDP San) and: 2-the !Xun and |Gui and || Gana contain appreciable amounts of Bantu-speaking (western African) ancestry (see sub-section 6.7) and western Africans have in general been well represented in genomic studies.

Allelic richness (number of distinct alleles in the population), private allelic richness (number of alleles private to the population) and the mean number of private alleles for pairs of populations were determined per base-pair using ADZE (Szpiech et al. 2008), applying the default settings. Results are plotted in Main Figure 1C, Main Figure 2 and Figure S5.5. The standard errors of the mean calculated by the ADZE were plotted in Main Figure 1C. The standard error of the mean for the allelic richness is of 0.005% for the Khoe-San, 0.008% for the other Africans and 0.007% for the non Africans. The corresponding values for the private allelic richness are 0.032%, 0.051% and 0.064%.

Allelic richness compares with heterozygosity estimates in that Khoe-San had the highest allelic richness (1.419 at the level of four individuals), even compared to a diverse group of four comparative Africans (1.401 at four individual level). Non-African allelic richness (1.285 at four individual level) were much lower compared to those of Africans. Within the Khoe-San allelic richness estimates also compares well with heterozygosity estimates with the !Xun having the highest richness and heterozygosity, followed by the |Gui and ||Gana, while the Ju|'hoansi had the lowest richness and heterozygosity. All of the Khoe-San groups on their own had higher allelic richness compared to western Africans (Mandenka and Yoruba) and much higher compared to Europeans and East Asians.

For alleles private to groups of individuals, Khoe-San had more private alleles compared to other Africans and non-Africans. Within the Khoe-San group, !Xun have the most private alleles and Ju|'hoansi the least. Most Khoe-San groups have similar amounts of private alleles but Ju|'hoansi seem to have observably less private alleles compared to the other groups. This could be due to a combination of appreciable outside admixture into the other groups (see sub-section 6.7), while the Ju|'hoansi is relatively un-admixed compared to other Khoe-San populations (see (Schlebusch et al. 2017)).

Alleles private to pairs of Khoe-San groups (Main Figure 1C), reflect Venn Diagram summaries, in that the two northern Khoe-San and the two southern Khoe-San groups share most alleles with each other, the central |Gui and ||Gana group shares intermediate number of alleles with all groups, and the least sharing are between any one northern and one southern Khoe-San groups. When outside groups are also considered in the private sharing of alleles (Main Figure 2), the Nama share comparatively more private alleles with non-Africans (Europeans); the !Xun and |Gui and ||Gana more with western Africans; the !Xun, |Gui and ||Gana and Nama with eastern Africans; and !Xun and Ju|'hoansi with Mbuti. Since the !Xun and Ju|'hoansi is geographically the most northern groups, their increased private allele sharing with a rainforest hunter-gatherer group is not unexpected. The central African Mbuti rainforest hunter-gatherer group seems to be a focal group, with high amounts of private allele sharing with most other African groups. The geographic area north of the Khoe-San groups and south of the rainforest hunter-gatherer groups, is today occupied by Bantu speakers who arrived relatively recently in the area (<5 kya (Li et al. 2014)), and replaced groups that possibly were genetically intermediate to the Khoe-San and central and eastern African hunter-gatherers. It is possible that the northern Khoe-San groups such as the !Xun had more gene flow with these extinct groups, and that we are observing signals of this gene flow in our analysis albeit with sub-optimal comparative groups. Also, it is possible that Bantu-speakers absorbed genetic components of these groups and subsequent admixture with the !Xun and |Gui and ||Gana introduced more of these components. The Ju|'hoansi and Nama individuals included in our study seem to have had the least Bantu-speaker admixture.

5.6. Site Frequency Spectrum Summary Statistics

Derived allele frequencies (DAF) for bi-allelic variants polymorphic in Khoe-San populations were extracted using VCFtools version 0.1.13 (Danecek et al. 2011). Estimates of mean site frequency spectrum (SFS) summary statistics (π, Tajima's *D*, Fay & Wu's *H*, DAF) for the whole genome and for protein coding sequence (exons) were computed using a weighted block jackknife approach (Busing et al. 1999) with a five megabase pairs genomic block removed for each iteration.

The Khoe-San groups exhibited relatively high nucleotide diversity overall (πKSP = 0.00098). This was especially apparent when some groups were compared to non-African populations (Naidoo et al. 2018). The Tajima's *D* estimates (Table S5.1) were also consistent with the populations' expected demographic histories, with negative values in overall estimates reflective of a population expansion. Protein coding sequence was found to be much less diverse (πKSP = 0.00052) than the whole genome average, as expected, due to high levels of purifying selection experienced by protein coding exons. This was corroborated by the Tajima's *D*, Fay & Wu's *H* and mean DAF estimates; though these estimates were strongly influenced by sample size. The Khoe-San groups, however, did not appear to differ substantially from each other based on these summary statistics. The extent of purifying selection on protein coding sequence and other non-coding genomic elements was examined in more detail in a related publication (Naidoo et al. 2018).

5.7. Small Indels

After mapping and quality procedures of the "KSP only dataset" (sections 2 and 4) including indel specific VQSR, we called a total of 2,176,524 small indels in the autosomes. Of these 1,267,661 were deletions, 908,863 insertions and 527,796 complex indels. The number of indels called in each Khoe-San population is given in Table S5.1.

5.8. Structural Variation

We used GenomeStrip (Handsaker et al. 2015) to call large scale deletions (relative to the human reference genome) among 23 of the 25 Khoe-San individuals specifically sequenced for this study. The two individuals with lower quality data were not included. Only the sequencing reslults (libraries) based on genomic DNA sequencing were included, libraries based on WGA DNA were not included in this analysis (see sub-section 2.4). The intention was not to generate a set of reliable copy-number variation (CNV) polymorphisms but to locate regions likely affected by structural variation in order to avoid these when searching for signals of selection. Hence, we only performed the 'preprocess' and 'discovery' part of the GenomeStrip pipeline (we did not perform any post-processing in refining the start and end positions of the regions or genotyping of the individuals and we did not match the regions to known CNVs). The method generated a set of regions were all the 23 individuals were called as having a deletion relative to the human reference genome. These are potentially interesting although we cannot at this point address whether the change has occurred on the Khoe-San branch or the reference genome or whether it represents an evolutionary event that resulted in the deletion of genetic material or a mutation that added genetic material to the genome. We identified 112 of these 'fixed deletions'. This set of deletions were further filtered to include only those that were separated by at least a 1000 bp of sequence to the closest structural variants called within the 1000 project (Auton et al. 2015) ending up with 44 CNVs so far only detected within our set of Khoe-San individuals. Genes separated by less than 50 kb of sequence for each of these deletions were recorded. Results are given in Table S5.3. Whether this set of genes were enriched for any GO-categories was investigated using DAVID (Huang, Brad T. Sherman, et al. 2009; Huang, Brad T Sherman, et al. 2009), see Table S5.4).

5.8.1. Evolutionary origin

Among the 44 Khoe-San specific deletions, all but three corresponded very well to a region that is missing among the three comparative ape genomes. For two (chr1:168024554-168025753 and chr5:143512864-143515050) of the three CNVs that did not follow this pattern, the sequence in the reference genome appears to be present in the three non-human apes. The third CNV

(chr16:46401931-46406443) is part of a much larger region missing in the apes and also very close to a large part of chromosome 16 that lacks reference sequence as well. Moreover, all individuals appear to have a maximum read coverage of 250 at all of the positions within this CNV.

In summary, although we are referring to these CNVs as deletions within the Khoe-San samples, they are more likely insertions on the non Khoe-San branch since theses regions are also missing in the non-human apes. At least, most of the evolutionary events appear to be on the genealogy of the reference genome and rare among Khoe-San individuals.

5.8.2. Function of genes close to fixed deletions

Among the 44 fixed CNVs, 24 (25 with *EGFR*) were within 50 kb of genes. Based on information from GeneCards (www.genecards.org) (Stelzer et al. 2016), among these CNVs, nine (ten with *EGFR*) are close to genes involved in skin or UV-protection, six are close to genes involved in immune response and five are close to genes associated with fat cells. All CNVs except one are located either inside an intron or in a non-coding region. The exception is chr6:56758337-56760961 which is located inside an exon of transcript variant 2 of the gene *DST* (a gene for which multiple alternatively spliced transcript variants encoding distinct isoforms have been found).

5.8.3. GO-analysis

There are only three GO-terms with an FDR<10 (Table S5.4), GO:0019557 (“histidine catabolic process to glutamate and formate”), GO:0019556 (“histidine catabolic process to glutamate and formamide”) and GO:0005604 (“basement membrane”). The first two of these are related to histidine metabolism. Interestingly, the deficiency of urocanic acid in histidine metabolism disorders could have implications for either or both of two proposed functions of urocanic acid - as a natural sunscreen against ultraviolet (UV) light and as a mediator of ultraviolet light-induced systemic immunosuppression (https://ommbid.mhmedical.com/content.aspx?bookid=971&sectionid=62674206&jumpsectionID=62674217). However, both the GO-terms associated with histidine metabolism are based on two genes close to the same fixed deletion (*AMDHD1*, *HAL*) and these do not have an FDR<10 in an additional (more subjective) analysis where only one gene is retained per fixed deletions. The third GO-term (which does have an FDR<10 for both analyses) is related to the collagen-containing extracellular matrix and as such potentially related to skin-structure.

5.9. Variant annotation

Variants were annotated using snpEff version 4.1 (Cingolani et al. 2012) with Ensembl database version 75 (GRCh37.p13) as the reference. Only canonical transcripts were used. Protein-coding region variants were further annotated with Combined Annotation-Dependent Depletion (CADD) (Kircher et al. 2014) to estimate relative functional significance. CADD is a general framework in which a metric of deleteriousness (the C score) is computed for any possible variant associated with a diverse range of annotations. The C score is based on a comparison between very high frequency variants in the human genome, and a simulated dataset of variants. While there would be a depletion of deleterious variants among very high frequency variants, due to purifying selection, there would be no such depletion in the simulated dataset. Additionally, Gowinda (Kofler and Schlötterer 2012) was used for gene set enrichment for potentially high impact variants.

While the vast majority of variants were found in intergenic or intronic regions of the genome, those variants associated with protein coding regions were also categorized based on their effect on the coding sequence (Table S5.5A). Across the 25 individuals, we discovered 62,364 missense variants. Within population counts of missense variants ranged from 30,857 and 31,298. There were fewer synonymous variants at 59,398 in total, however, each group appeared to have slightly more synonymous than missense variants (31,844 to 32,911). The higher numbers of synonymous variants per population, and the greater overlap in synonymous variants among the populations were both consistent with higher levels of purifying selection acting on missense variants. We also focused on a subset of potential loss-of-function (LOF) variants, specifically those that occurred due to a single base substitution. While missense variants have the potential to inactivate a gene, we did

not include them as LOF variants; due to the diversity of their impact. In total, the LOF variants included 994 stop-gain variants, 758 splice donor variants, 666 splice acceptor variants, 86 start-loss variants, and 70 stop-loss variants. Individual genomes were also annotated (Table S5.6). With regard to LOF variants, the average individual for the total sample contained 121.5 (SD: 19.2) stop-gain variants, 187.4 (SD: 13.1) splice donor variants, 140.4 (SD: 10.1) splice acceptor variants, 19.2 (SD: 3.3) start-loss variants, and 19.4 (SD: 2.4) stop-loss variants.

5.9.1. Estimation of functional significance

Protein coding region variants from our dataset were annotated with CADD (Kircher et al. 2014) in order to estimate the functional significance of each variant. (Kircher et al. 2014) calculated a C score for all 8.6 billion possible single nucleotide variants (SNVs) of the GRCh37 reference. They then defined a scaled C score, on the basis of the rank of each variant, which ranged from 1 to 99. Variants in the highest 10% of raw C scores, for example, would be assigned scaled C scores 10 and greater, while variants in the highest 1% of raw C scores would be assigned scaled C scores 20 and greater. We calculated the proportion of each of the variant types per scaled C score (Figure S5.6A). The highest scores were associated almost exclusively with stop-gain variants, with over 96% of them scoring 20 and above, i.e. the vast majority of stop-gain variants found were among the top 1% of functionally significant variants. This result was reflected by a much lower mean allele frequency for stop-gain variants (Table S5.5B), due to the effects of purifying selection. While the bulk of the distributions of the other coding region variants fell below a scaled C score of 20, approximately one third of missense, start-loss, splice acceptor and splice donor variants scored 20 and greater (Figure S5.6B). Unsurprisingly, only 2% of synonymous variants scored 20 and greater, while only 8.6% of stop-loss variants scored 20 and greater, indicating that this class may be the least deleterious LOF variant.

5.9.2. Biological roles

We performed a gene ontology (GO) enrichment analysis on LOF variants and potentially high impact missense variants (scaled C score $\geq$ 20) to elucidate biological functions that may be affected by these variants. Stop-gain variants were associated primarily with the detection of chemical stimuli (smell and taste), immune response, and toxin metabolism in terms of biological processes; while intermediate filaments (including keratin filaments) were primarily affected with regard to cellular components. Start-loss and stop-loss variants had fewer significant hits and were primarily associated with the detection of chemical stimuli. Splice acceptor and donor variants had no significant hits. Potentially high impact missense variants, due to their sheer number (20,238), were associated with a multitude of GO terms including some of those mentioned above. Detection of stimulus was a major theme with regard to biological processes, though this was represented by a broad range of stimuli. Other notable terms included RNA surveillance, amino acid activation, cofactor transport and microtubule cytoskeleton organization. With regard to cellular components, intermediate filaments were also enriched for missense variants (see Table S5.7 for full list).

We considered the moderate to high frequency LOF variants in more detail, screening start-loss and stop-loss variants above 10% (Tables S5.8 and S5.9), and stop-gain variants above 40% (Table S5.10). Notably, almost all LOF variants examined in detail were previously discovered. Most of the LOF variants present in the Khoe-San were found at frequencies comparable to the global average, or to other African populations (Auton et al. 2015). They were also usually found in genes with multiple transcripts, and so were more likely to be partial LOF variants at most (MacArthur et al. 2012). There were, however, a few instances of variants that occurred at noticeably differing frequencies to other populations, or were associated with noteworthy consequences.

5.9.3. Notable common LOF variants in the Khoe-San

Start-loss variants

rs17124277: a G-to-A variant in the *BPIFA2* gene. This variant, which has been found at frequencies of 7-16% in African populations, and seldom found outside of Africa (Auton et al. 2015), was present in the Khoe-San at 36%. *BPIFA2* encodes the Parotid Secretory Protein (PSP), a

member of the Plunc (palate, lung and nasal epithelium clone) family of proteins (Prokopovic et al. 2014). PSP is a soluble salivary protein which acts as part of the innate immune response. It exhibits anti-bacterial activity, particularly against *Pseudomonas aeruginosa (Geetha et al. 2003)*.

rs113094669: a G-to-A variant in the *OR6Q1* gene. This variant, which has only been found in one African American individual (Auton et al. 2015), was present in the Khoe-San at 32%. *OR6Q1* encodes the Olfactory receptor 6Q1. While several olfactory receptor genes were found to contain LOF variants, the switching off of this gene was notably specific to the Khoe-San.

rs141151195: a T-to-C variant in the *SPDYC* gene. This variant, which has been found at a frequency of 3% in the Luhya (Auton et al. 2015) and sporadically in other cohorts (Fu et al. 2012; Lek et al. 2016), was present in the Khoe-San at 14%. *SPDYC* encodes Speedy protein C. This protein is involved in cell cycle progression, through activation of of cyclin-dependant kinases (CDK1 and CDK2) (Cheng and Solomon 2008).

rs113389918: a T-to-C variant in the *SFRP4* gene (may be partial LOF). This variant, which has been found only in a few individuals (one in the NHLBI Exome Sequencing Project (ESP6500) cohort (Fu et al. 2012), one in the Exome Aggregation Consortium (ExAC) cohort (Lek et al. 2016), and in two Khoe-San individuals (Schuster et al. 2010)), was present in the Khoe-San at 14%. *SFRP4* encodes Secreted frizzled-related protein 4 which acts as a modulator of Wnt signalling pathways (Mahdi et al. 2012). These are highly conserved signal transduction pathways that regulate processes such as gene transcription and cytoskeleton formation (Nusse 2005). (Mahdi et al. 2012) showed that high levels of expression of *SFRP4* were linked to the onset of Type 2 Diabetes a few years later. A reduction in SFRP4 was shown to affect the functioning of islet cells, thus, decreased expression of *SFRP4* could possibly be protective against Type 2 Diabetes.

rs146195386: a A-to-G variant in the *IFNA13* gene. This variant, which has been found at very low frequencies (0.2% in ExAC (Lek et al. 2016)), was present in the Khoe-San at 14%. *IFNA13* encodes the Interferon alpha 13 protein; a member of the Interferon family of cytokines, which display antiviral, anti-proliferative, and immunomodulatory effects (Lengyel 1982). It should be noted that the protein sequence of Interferon alpha 13 is identical to that of Interferon alpha 1, and so this may mitigate any losses to alpha 13.

rs112330886: a T-to-G variant in the *PRAMEF1* gene (may be partial LOF). This variant, which has been found at very low frequencies (0.1% in ExAC (Lek et al. 2016)), was present in the Khoe-San at 10%. *PRAMEF1* is a member of the Preferentially expressed antigen of melanoma (PRAME) family, which are expressed in testicular tissue and in tumors such as melanoma. These proteins may be involved in the positive regulation of cell proliferation, and in negative regulation of apoptosis. The PRAME genes are located on a copy number polymorphic region on chromosome 1, and have evolved as a family through numerous duplications (Birtle et al. 2005). Some of these genes have also been pseudogenized.

rs543701166: a C-to-T variant in the *TMEM207* gene. This variant was present in the Khoe-San at 10%, with no other frequency information available for other populations. *TMEM207* encodes Transmembrane protein 207. This protein is not well characterised, however, it has been found that expression of TMEM207 was able to increase the invasive activity of KATO-III gastric carcinoma cells, *in vitro* (Takeuchi et al. 2012). This may be accomplished through its binding of the tumor-suppressor WW domain-containing oxidoreductase (WWOX). Since *TMEM207* appeared to be highly expressed in several gastric signet-ring cell carcinoma tissue specimens, it may be an oncogene.

Stop-gain variants

rs497116: a G-to-A variant in the *CASP12* gene. This well characterised stop-gain variant is fixed in most human populations, and results in a non-functional Caspase-12 protein. The functional form, however, has been found in African populations at frequencies of 14-23% (Auton et al. 2015). Notably, it appears at a frequency of 48% in the Khoe-San. The functional Caspase-12 has been associated with increased risk of sepsis, as it is involved in the down-regulation of inflammatory

cytokines (Saleh et al. 2004; Saleh et al. 2006). Studies have examined why the functional allele persists in some populations ; investigating links to candidemia (Rosentul et al. 2012), rheumatoid arthritis (Marshall et al. 2014), and obesity (Skeldon et al. 2016).

rs17147990: a T-to-A variant in the *HTN3* gene. This variant, which has been found at frequencies of 12-21% in African populations, and seldom found outside of Africa (Auton et al. 2015), was present in the Khoe-San at 50%. *HTN3* encodes the Histatin 3 protein. Histatins are salivary proteins with antimicrobial and antifungal properties (Troxler et al. 1990), and function as part of the innate immune system. They also appear to have a role in wound healing (Oudhoff et al. 2008). The stop-gain variant occurs in the fifth exon of the gene, and so any effect may be mitigated by its presence close to the end of the protein.

rs3213755: a G-to-A variant in the *KRTAP1-1* gene. This variant, which has been found at frequencies of 15-19% (ESP6500, ExAC), was present in the Khoe-San at 42%. *KRTAP1-1* encodes the keratin-associated protein 1-1. Keratin-associated proteins, together with keratins, are the main structural components of hair (Shimomura et al. 2003).

rs6661174: a C-to-T variant in the *FMO2* gene. This nonsense variant (similar to rs497116) is almost fixed in most human populations, and results in a non-functional Flavin Containing Monooxygenase 2 protein. The functional form, however, has been found in African populations at frequencies of 12-18% (Auton et al. 2015). Notably, it appears at a frequency of 60% in the Khoe-San. The functional FMO2 protein is able to metabolise thiourea; however, in doing so, produces more toxic derivatives (Veeramah et al. 2008). Carriers of the functional allele are at increased risk for pulmonary toxicity when exposed to thiourea, which is present in a wide range of industrial, household and medical products.

Stop-loss variants

rs200329830: a T-to-G variant in the *CGB1* gene. This variant, which has been found at frequencies of 1.2% (ExAC), was present in the Khoe-San at 34%. *CGB1* encodes Chorionic Gonadotropin Beta Subunit 1, which is part of the Gonadotropin family of polypeptide hormones. The Gonadotropins are central to the endocrine system, and serve to regulate the functions of growth, sexual development and reproduction. The functions of *CGB1* and *CGB2*, however, are unclear; as these loci produce novel proteins due to frameshift variants (Hallast et al. 2007).

rs7276273: an A-to-C variant in the *KRTAP10-4* gene. This variant, which has been found at frequencies of 19-35% in African populations, and found only at low levels (0-6%) outside of Africa (Auton et al. 2015), was present in the Khoe-San at 34%. *KRTAP10-4* encodes the keratin-associated protein 10-4. Keratin-associated proteins, together with keratins, are the main structural components of hair (Shimomura et al. 2003).

rs61894893: a T-to-C variant in the *TRIM48* gene. This variant, which has been found at frequencies of 0-10% across the world (Auton et al. 2015), was present in the Khoe-San at 30%. *TRIM48* encodes the Tripartite Motif Containing 48 protein, a member of the Tripartite motif (TRIM) family. These proteins are part of the innate immune system, and are involved in recognition of pathogens as well as regulating host defence transcriptional pathways.

rs28375936: a T-to-C variant in the *NPIPB6* gene. While no population frequency information is available for the variant, it has been found as a somatic mutation in thyroid tumour tissue (COSMIC project (Forbes et al. 2017)). It was found in the Khoe-San at a frequency of 12%. *NPIPB6* encodes the Nuclear Pore Complex Interacting Protein Family Member B6. While it is unclear what these proteins do, the family appears to be well conserved in primates, including humans (Johnson et al. 2001).

5.9.4. Discussion

Following the high coverage whole genome sequencing of 25 Khoe-San individuals, we extracted a dataset of variants associated with human protein coding genes. The average estimates of LOF variants per genome were substantially higher than those found in some previous studies (Auton et

al. 2015; Mallick et al. 2016). It is, however, quite difficult to compare estimates obtained in different studies, as these are dependent on several factors. Variants examined as LOF often differ between studies. In our case, we did not include frame-shift mutations. It should also be noted that LOF variants show enrichment for sequencing and annotation errors (MacArthur and Tyler-Smith 2010; MacArthur et al. 2012); and though a stringent filtering protocol was followed when calling these variants, care should always be taken when assigning clinical or functional relevance to any potential LOF variants. Other studies that utilized a more aggressive filtering protocol (MacArthur et al. 2012) or those that flagged variants unlikely to result in LOF (Auton et al. 2015) obtained lower estimates than ours. Since high coverage sequence was generated for this study, this allowed for greater sensitivity in discovering singletons (which are enriched among LOF variants) compared to studies that made use low coverage sequence data. This also mitigated, to some extent, the level of sequence error. The sampled population also contributes to the number of LOF variants found. The 1000 Genomes Project (Auton et al. 2015) found that African genomes contained ~30 more filtered LOFs than non-African populations. With the higher diversity found in Khoe-San populations, our higher numbers appear consistent. Finally, choice of transcript and gene annotation, and effect predicting software will also influence LOF numbers (McCarthy et al. 2014).

With possibly the exception of stop-loss variants, we found that a large proportion of the LOF variants in our dataset scored highly in estimations of functional significance, especially compared to synonymous variants. The interpretation of the presence of a LOF variant, however, is quite difficult since the disruption of a protein coding gene may lead to numerous possible effects. The loss of gene function may cause genetic diseases such as Cystic fibrosis or Spinal muscular atrophy; resulting in reduced quality of life and mortality. Yet high numbers of LOF variants have also been found in seemingly healthy individuals (Yngvadottir et al. 2009; Auton et al. 2015), indicating that not all LOF variants are likely to be pathogenic.

While a large proportion of the LOF variants found in our sample were singletons, potentially under selective constraint, the more common LOF variants in the Khoe-San were often found in genes with closely related paralogs. These genes were also often associated with multiple transcript isoforms. This was reflective of previous findings regarding LOF variants in human genomes (MacArthur et al. 2012). This presence of LOF variants mainly in potentially redundant genes is also congruent with the removal of deleterious forms from the gene pool via purifying selection.

The large number of genes associated with detection of chemical stimuli, specifically smell (olfaction), with LOF variants was not surprising, as olfactory receptors constitute the largest gene family in mammals; however, in humans around 60% of them are pseudogenes (Menashe et al. 2003). An increased rate of pseudogenization seemed to be specific to humans compared to other primates (Gilad et al. 2003), which may be suggestive of lower levels of purifying selection on the gene family. Moreover, several of these are polymorphic pseudogenes, and so both functional and pseudogenic alleles segregate in human populations. Thus most human populations have very diverse repertoires of working olfactory receptors. *OR6Q1* may be an example of one of these, though it appears to be polymorphic only in the Khoe-San. The keratin and keratin-associated proteins are also examples of large gene families which showed enrichment for LOF variants. The LOF variants in the keratin-associated proteins, *KRTAP1-1* and *KRTAP10-4*, appeared to be much more common in African populations. Keratin-associated proteins have previously been shown to contribute toward hair variation in mammals (Khan et al. 2014), and so moderate to high frequency disruptive polymorphisms such as those found in *KRTAP1-1* and *KRTAP10-4* should be investigated for their potential to contribute to hair structure variation in human populations.

Functional disruptions in immune response genes are often caused by rare variants, and may contribute to the burden of disease in a population (Rausell et al. 2014) especially if they occur in immune genes under high selective pressure. They may also contribute to inter-individual variability of immune response. LOF variants in the immune genes *BPIFA2*, *IFNA13*, *HTN3*, and *TRIM48* genes in the Khoe-San, however, occurred at moderate to high frequencies; and so may point to higher levels of redundancy in these gene families, or possibly toward localized reductions in purifying selection. The high frequencies of functional *CASP12* and *FMO2* genes in the Khoe-

San was also quite intriguing. The functional form of the *CASP12* gene is found at a frequency of 48% in our sample, while the global average is around 5%. Functional Caspase-12 protein has been associated with an increased risk of sepsis, as it is involved in the down-regulation of inflammatory cytokines (Saleh et al. 2004; Saleh et al. 2006). Likewise, the functional form of the *FMO2* gene is found in our sample at a frequency of 60%. The functional Flavin-containing Monooxygenase 2 protein is able to metabolise thiourea; however, in doing so produces toxic derivatives (Veeramah et al. 2008). Carriers of the functional allele are at increased risk for pulmonary toxicity when exposed to thiourea, which is present in a wide range of industrial, household and medical products. This may possibly point to differing selective pressures experienced by these populations.

5.10. mtDNA and Y-chromosome

5.10.1. mtDNA

Methods

We analysed the mitochondrial genome of the Khoe-San and most of the other Africans samples from the "global" dataset (see details of the samples in sub-sections 3.2, 3.3 and 3.4). For most populations five individuals were used for the analysis (one population had only four individuals). In total, we analysed 59 individuals from twelve populations representing hunter-gatherers from western, southern and eastern Africa and farmers from western and eastern Africa.

Variant calling and filtering

We used the UnifiedGenotyper module of GATK (McKenna et al. 2010) to call mtDNA SNPs from BAM files for the KSP samples according to GATK Best Practices recommendations (DePristo et al. 2011; Van der Auwera et al. 2013). We did not consider indels, the default SNP genotype likelihoods calculation mode was used and the ploidy argument set to 1. Both the all sites (containing invariant sites) and the variant sites VCFs were generated for downstream analysis.

We applied different filtering methods on the raw SNPs depending on the sample source. For the KSP samples, raw variants were filtered using GATK's tool VariantFiltration. The QualByDepth, Depth, MappingQuality and FisherStrand annotations were extracted to determine the hard-filtering threshold, so that 5% of the lowest quality records were filtered out. For comparative data (selected African individuals from KGP (Auton et al. 2015), CG (Drmanac et al. 2010) and LC (Lachance et al. 2012) datasets), variants were filtered using in-house scripts with the following parameters: minimum depth 50 X, minimum genotype quality 50, and excluding records that were marked "VQLOW".

We merged the filtered SNP call sets from KSP, KGP, CG and LC using the vcf-merge function from the VCFtools package (Danecek et al. 2011), and removed sites with > 5% missing calls. Moreover, we excluded two poly-C regions (np 303-315, 16183-16192) from the analysis.

Haplogroup assignments

We used the online tool Haplofind (Vianello et al. 2013) to assign haplogroups for mtDNA sequences. Haplofind classifies complete mtDNA sequences according to haplogroup nomenclature based on PhyloTree (van Oven and Kayser 2009).

Phylogenetic analysis

We reconstructed phylogenetic trees and dated various nodes using BEAST v1.8.2 (Drummond et al. 2012). VCF files of complete mtDNA genomes were converted to FASTA files using Bergey's (2012) Perl scripts vcf-tab-to-fasta (http://code.google.com/p/vcf-tab-to-fasta), which were then imported to BEAUti (a graphical user-interface application included in BEAST package) to generate BEAST input XML files.

We chose the best-fitting substitution models by conducting test runs in jModelTest v2.7.1 (Darriba et al. 2012). The tree model was set to Coalescent: Bayesian Skyline (Drummond et al. 2005) with a

piecewise-linear skyline model. We selected the HKY+I+G substitution model, with a strict clock model and a mutation rate of 1.665 x10-8 /bp/year (Soares et al. 2009).

We performed multiple runs for the dataset, using 50 to 100 million MCMC iterations, with a sampling every 1,000th step. The initial 10% of each run was discarded as burn-in and the outputs were inspected in Tracer v1.6 (Rambaut et al. 2018) and confirmed that all EES values were above 200.

We annotated Maximum clade credibility (MCC) trees in TreeAnnotator (Drummond et al. 2012) and extracted the mean, median, and 95% HPD intervals of the node heights for dating. The trees were visualized and edited in TreeGraph v2.7.1 (Stöver and Müller 2010) for displaying.

Results

After filtering and cleaning the data for our 25 Khoe-San samples, we found 316 mtDNA SNPs. For the combined dataset we found 511 mtDNA SNPs.

mtDNA haplogroups

Table S5.11 shows the 25 Khoe-San individuals that we sequenced together with their populations, their language affiliation and their mtDNA haplogroups. Our results showed a high frequency of L0d with 20 of the 25 individuals harboring this deep rooting macro-haplogroup. Out of the five remaining individuals, three have haplogroup L0k, another deep-rooting but geographically more restricted haplogroup. Only two individuals have other haplogroups than L0d and L0k (which are the “typical Khoe-San associated haplogroups”). One of these individuals has L1c and the other has L4.

Due to our small sample size, we compared our results with haplogroup frequencies obtained from previous studies on the same (or similar) hunter-gatherer groups with larger sample sizes. Table S5.12 shows the frequency distribution of haplogroup L0d, L0k, L1 and L4. Combined with the results in Table S5.11, we can estimate whether our sample has a common occurring or rare haplogroup in terms of the population the sample belongs to.

L0d is the most prevalent haplogroup overall in Khoe-San groups, ranging in frequency from 55% in the !Xun to 100% in the Karretjie People (Schlebusch et al. 2011). Thus it is not unexpected that 80% of our samples carry L0d and most of our samples carry a common sub-haplogroup.

We can also compare frequencies at a population level. Among our five !Xun individuals, three of them carry L0k, the other two carry L0d1. The corresponding frequency of L0k and L0d1 in a larger sample of !Xun (Barbieri et al. 2014) are 33.3% and 44.4%. Similarly, three |Gui and ||Gana and four Ju|'hoansi individuals carry L0d1, which is the most common haplogroup in these two populations (60.9% and 50.0%). L0d2 reaches its highest frequency in Karretjie people (58.1%), and in our samples four out of the five Karretjie individuals carry L0d2. L0d2 is also one of the most common haplogroups in the Nama (34.5%, while the most common haplogroup, L0d1, has a frequency of 37.9%), and was found in four of our Nama individuals.

Interestingly, we found one |Gui and ||Gana individual carrying L1c. L1c predominates in central Africa (Quintana-Murci et al. 2008), with highest frequencies (77% to 100%) in the rain-forest hunter-gatherers, in particular L1c1a (Tishkoff et al. 2007). The other L1c sub-haplogroups display variable frequencies in Bantu-speaking agricultural populations (Batini et al. 2007). In general, a higher frequency of L1c in central Africa and the decreased frequencies in western and southeastern Africa indicate an origin or early arrival of L1c in central Africa. Our |Gui and ||Gana sample has L1c2, which is not the rain-forest hunter-gatherer-specific sub-haplogroup, suggesting it came from Bantu-speakers agriculturalists who brought L1c to southern Africa. Most likely, it is an indication of Bantu-speakers admixture into Khoe-San populations, rather than gene flow between central and southern hunter-gatherers.

We also found one Ju|'hoansi individual carrying L4, a rare African haplogroup most frequent in eastern and northeastern Africa (Salas et al. 2002; Kivisild et al. 2004). The highest frequencies are

in Tanzania among Hadza (60%) and Sandawe (48%) (Tishkoff et al. 2007). L4 is uncommon in Khoe-San populations and has been only reported by (Barbieri et al. 2014) where three Ju|'hoansi carried this haplogroup. Our L4 individual comes from the same population.

mtDNA tree

Figure S5.7 shows the mtDNA phylogenetic tree of the combined dataset that we obtained using BEAST. The topology of our mtDNA tree matches published trees that focused on the deepest-rooting lineages of mtDNA phylogeny (Behar et al. 2008; Barbieri et al. 2013; Schlebusch et al. 2013). Different sub-lineages can be distinguished from the topology of the tree, and in agreement with the haplogroup assignment (done using HaploFind), individuals assigned to have the same sub-haplogroup are unambiguously grouped together on the tree. We dated the tree using a Bayesian approach implemented in BEAST. The tree coalesces 180 kya (95% C.I.: 155-204 kya), when L0 separates from the rest (L1-4), and splits into L0d and L0abk. The coalescent time of L0d and L0k is 145 kya (95% C.I.: 122-170 kya), consistent with the previous dating estimates (Barbieri et al. 2013). All the individuals belonging to L0k and L0d clades are exclusively from the newly sequenced Khoe-San populations in our study. The L4 sequence from our study (found in one Ju|'hoansi individual) groups with two Hadza and one Maasai from eastern Africa. As mentioned before for the one L1c sequence in a |Gui and ||Gana individual, this L1c sequence does not belong to the rain-forest hunter-gatherer-specific L1c1a sub-haplogroup but rather to L1c2. Correspondingly on the tree all rain-forest hunter-gatherer individuals form a L1c1a clade, which leaves the L1c2 |Gui and ||Gana individual as a more distantly related sister group.

5.10.2. Y chromosome

The Y chromosomes of the 19 male individuals in our sample were analysed in (Naidoo et al. In revision). The haplogroups assigned with AMY-tree v.2.0 (Van Geystelen et al. 2013) are reported in Table S5.11. The major haplogroups found were A-M14 (A1b1a1), A-M51 (A1b1b2a), B-M112 (B2b), E-M2 (E1b1a1) and E-M35 (E1b1b1).

5.10.3. Shared ancestry between southern and eastern African hunter-gatherers

Hadza and Sandawe are (or were) hunter-gatherers who also speak languages with click consonants. It has been hypothesized that the eastern African Hadza and Sandawe derive a fraction of their ancestry from a Khoe-San-related hunter-gatherer population who once occupied a wide region of southern and eastern Africa (Tobias 1964). Whole genome studies using autosome data have found evidence for gene flow between Khoe-San and eastern Africans (especially between Ju|'hoansi and Hadza) (Pickrell et al. 2012; Schlebusch et al. 2012; Skoglund et al. 2017) - see also sub-section 6.7. Here we detected haplogroup sharing between Ju|'hoansi and Hadza in both mtDNA and Y chromosome data. Interestingly, the TMRCA (time to the most recent ancestor) for Ju|'hoansi and Hadza in mtDNA haplogroup L4 and Y chromosome haplogroup B2b are very close (36 kya (95% C.I.: 25-50 kya), and 33 kya (95% C.I.: 24-43 kya), respectively).

Variant summaries Tables and Figures

Table S5.1: Autosomal variant summaries

AUT SNP files	Total	Karretjie	Nama	|Gui and ||Gana	Ju|'hoansi	!Xun
SNP aut (unfiltered)	20020719	11381060	11352483	11468417	11267114	11490865
SNP aut (VQSRed)	19001637	10811200	10766197	10911185	10680611	10937095
Dinucleotide SNPs (filtered miss, HWE)	18637959	10555587	10514246	10649570	10429573	10676563
Exonic SNPs (%)	0.653	0.597	0.598	0.596	0.599	0.598
Ave nr SNPs per ind (SDev)	5,079,953 (46,214)	5094845	5022536	5072602	5118469	5091313
Novel variants vs dbSNP built 151 (% of variants)	1,960,665 (10.5%)	578,935 (5.5%)	632,492 (6.0%)	491,543 (4.6%)	548,528 (5.3%)	477,360 (4.5%)
Singletons (% of variants)	5,403,107 (29.0%)	4,547,752 (43.1)	4,504,833 (42.8)	4,639,215 (43.6)	4,315,691 (41.4)	4,660,181 (43.6)
Non-singleton novel variants (% of variants)	602,402 (3.2%)	108,129 (1.0%)	95,736 (0.9%)	97,282 (0.9%)	106,546 (1.0%)	91,365 (0.9%)
Mean Depth, duplicates excluded (All pos in ref genome)	53.4 (45.1 – 59.8)	52.5 (45.1 – 56.9)	55.3 (53.7 – 56.4)	51.7 (48.2 – 54.8)	52.8 (48.0 – 59.8)	54.7 (51.7 – 57.1)
Mean Depth (Called+Filtered Vars)	52.1 (44.5-58.2)	51.0 (44.5-53.6)	53.5 (50.2-55.8)	50.5 (46.7-54.0)	51.4 (46.7-58.2)	54.4 (51.5-56.7)
SNP Missingness (Called+Filtered Vars)	0.072% (0.05%-0.17%)	0.081%	0.082%	0.070%	0.106%	0.060%
%Uncalled or missing compared to reference	1.278% (1.209%-1.343%)	1.276%	1.277%	1.276%	1.302%	1.256%
Mean Heterozygosity (Genomic)	0.001274	0.001273	0.001266	0.001275	0.001263	0.001291
Heterozygosity (Called+Filtered Vars)	0.183249	0.183205	0.182149	0.183438	0.181665	0.185790
Tajima's D	-0.783	-0.335	-0.320	-0.361	-0.283	-0.364
Tajima's D (exonic)	-1.141	-0.520	-0.484	-0.548	-0.473	-0.545
π	0.000983	0.000965	0.000963	0.000972	0.000953	0.000973
π (exonic)	0.000516	0.000505	0.000507	0.000509	0.000501	0.000511
FayWu H	-0.000503	0.024947	0.0259495	0.0271395	0.023198	0.027285
FayWu H (exonic)	0.013394	0.047238	0.048597	0.048635	0.045592	0.049412
Mean DAF	0.175	0.275	0.275	0.273	0.277	0.272

Mean DAF Exonic	0.152	0.258	0.260	0.257	0.261	0.257
Total Indels (VQSRed)	2176524	1441604	1458457	1433307	1439949	1461609
Deletions	1267661	802507	799461	812743	795648	815037
Insertions	908863	634150	634609	640933	631300	642247
Complex indel	527796	513400	512696	514178	512684	514099
Structural Variants	4452	1979	2030	2419	2139	2362
Proportion Structural variants with genes	0.378	0.39	0.37	0.379	0.377	0.374

Table S5.2: Concordance rate of the callset (SNPs) and average coverage for the 25 Khoe-San samples.

KSP	Concordance rate	Average coverage (duplicates removed)	Stdev coverage (duplicates removed)	Average coverage (duplicates included)
KSP062	0.9913	53.18	88.24	54.06
KSP063	0.9681	56.91	166.69	59.71
KSP065	0.9906	56.39	80.99	57.59
KSP067	0.9915	45.09	83.66	47.96
KSP069	0.9914	51.12	80.40	54.09
KSP092	0.9912	54.75	79.13	56.46
KSP096	0.9888	48.17	204.33	50.43
KSP103	0.9806	48.40	85.64	49.71
KSP105	0.9897	47.96	95.04	51.61
KSP106	0.9880	59.77	96.85	61.30
KSP111	0.9912	53.26	98.53	55.36
KSP116	0.9913	54.48	78.12	55.67
KSP124	0.9715	54.28	277.50	58.07
KSP134	0.9909	55.76	83.83	56.97
KSP137	0.9906	53.71	103.48	55.03
KSP139	0.9919	56.38	89.36	57.68
KSP140	0.9916	56.42	88.99	57.68
KSP146	0.9915	55.63	82.35	56.78
KSP150	0.9897	53.79	87.84	55.30
KSP152	0.9896	57.07	82.65	58.52
KSP154	0.9913	51.71	72.29	52.56
KSP155	0.9909	55.36	82.02	56.39
KSP224	0.9917	53.20	103.55	54.72
KSP225	0.9917	52.89	94.25	54.17
KSP228	0.9914	49.60	80.62	51.28
AVERAGE	**0.9887**	**53.4111**	**102.653**	**55.1632**
MINIMUM	**0.9681**	**45.0893**	**72.2908**	**47.9598**
MAXIMUM	**0.9919**	**59.7655**	**277.5038**	**61.2991**

Table S5.3: Detail of 44 “fixed deletions” in 23 Khoe-San samples >1000 bp away from any structural variants called within the 1000 genomes project. Notes are from genecards.org (Stelzer et al. 2016).

Region	Genes overlapping 50 kb flanks	Notes
Chr1:83125953-83127590	-	-
Chr1:84517925-84524650	PRKACB	PRKACB (Protein Kinase CAMP-Activated Catalytic Subunit Beta): Diseases associated with PRKACB include Primary Pigmented Nodular Adrenocortical Disease and Carney Complex Variant. Among its related pathways are Melanocyte Development and Pigmentation and Ovarian steroidogenesis. Involved in the regulation of lipid and glucose metabolism and is a component of the signal transduction mechanism of certain GPCRs.
Chr1:156526706-156528955	IQGAP3, TTC24	IQGAP3 (IQ Motif Containing GTPase Activating Protein 3): Among its related pathways are Signaling by Rho GTPases and RHO GTPases activate IQGAPs. TTC24 (Tetratricopeptide Repeat Domain 24).
Chr1:168024554-168025753	DCAF6, GPR161	DCAF6 (DDB1 And CUL4 Associated Factor 6): The protein encoded by this gene is a ligand-dependent coactivator of nuclear receptors, including nuclear receptor subfamily 3 group C member 1 (NR3C1), glucocorticoid receptor (GR), and androgen receptor (AR). The encoded protein and DNA damage binding protein 2 (DDB2) may act as tumor promoters and tumor suppressors, respectively, by regulating the level of androgen receptor in prostate tissues. In addition, this protein can act with glucocorticoid receptor to promote human papillomavirus gene expression. GPR161 (G Protein-Coupled Receptor 161): Diseases associated with GPR161 include Pituitary Stalk Interruption Syndrome and Spindle Cell Hemangioma.
Chr1:184814724-184820800	FAM129A	FAM129A (Family With Sequence Similarity 129 Member A).
Chr1:197500501-197503285	DENND1B	DENND1B (DENN Domain Containing 1B): Diseases associated with DENND1B include Asthma and Interleukin-7 Receptor Alpha Deficiency. Guanine nucleotide exchange factor (GEF) for RAB35 that acts as a regulator of T-cell receptor (TCR) internalization in TH2 cells.
Chr1:236549084-236551345	EDARADD	EDARADD (EDAR Associated Death Domain): This gene was identified by its association with ectodermal dysplasia, a genetic disorder characterized by defective development of hair, teeth, and eccrine sweat glands. The protein encoded by this gene is a death domain-containing protein, and is found to interact with EDAR, a death domain receptor known to be required for the development of hair, teeth and other ectodermal derivatives. This protein and EDAR are coexpressed in epithelial cells during the formation of hair follicles and teeth. Through its interaction with EDAR, this protein acts as an adaptor, and links the receptor to downstream signaling

		pathways. Two alternatively spliced transcript variants of this gene encoding distinct isoforms have been reported.
Chr2:208474482-208476801	METTL21A,CREB1	METTL21A (Methyltransferase Like 21A): Among its related pathways are Metabolism of proteins and Protein methylation. CREB1 (CAMP Responsive Element Binding Protein 1): Diseases associated with CREB1 include Histiocytoma, Angiomatoid Fibrous and Melanoma Of Soft Tissue. Involved in different cellular processes including the synchronization of circadian rhythmicity and the differentiation of adipose cells.
Chr2:217089934-217092507	XRCC5, MARCH4	XRCC5 (X-Ray Repair Cross Complementing 5): A rare microsatellite polymorphism in this gene is associated with cancer in patients of varying radiosensitivity. Diseases associated with XRCC5 include Werner Syndrome [premature aging with increased risk of malignant melanoma] and Alpha-Thalassemia/ Mental Retardation Syndrome, X-Linked. Among its related pathways are HIV Life Cycle and Coregulation of Androgen receptor activity. Involved in DNA non-homologous end joining (NHEJ) required for double-strand break repair and V(D)J recombination. Plays a role in the regulation of DNA virus-mediated innate immune response by assembling into the HDP-RNP complex, a complex that serves as a platform for IRF3 phosphorylation and subsequent innate immune response activation through the cGAS-STING pathway. MARCH4 (Membrane Associated Ring-CH-Type Finger 4): reduces surface accumulation of several membrane glycoproteins by directing them to the endosomal compartment.
Chr3:83853112-83855362	-	-
Chr3:137072351-137073461	-	-
Chr3:152311717-152313171	-	-
Chr4:46055989-46058214	GABRG1	GABRG1 (Gamma-Aminobutyric Acid Type A Receptor Gamma1 Subunit): Diseases associated with GABRG1 include Alcohol Dependence and Autism. Among its related pathways are Akt Signaling and GABAergic synapse. GABA, the major inhibitory neurotransmitter in the vertebrate brain.
Chr4:79269116-79275213	FRAS1	FRAS1 (Fraser Extracellular Matrix Complex Subunit 1): diseases associated with FRAS1 include Fraser Syndrome 1 and Renal Hypodysplasia/Aplasia 3. This gene encodes an extracellular matrix protein that appears to function in the regulation of epidermal-basement membrane adhesion and organogenesis during development. Mutations in this gene cause Fraser syndrome, a multisystem malformation that can include craniofacial, urogenital and respiratory system abnormalities. Alternative splicing results in multiple transcript variants.
Chr4:108127809-108131736	DKK2	DKK2 (Dickkopf WNT Signaling Pathway Inhibitor 2): Among its related pathways are Reelin Pathway (Cajal-Retzius cells) and HIV Life Cycle. DKKs play an important role in vertebrate development, where they locally inhibit Wnt regulated processes such as antero-posterior axial patterning, limb development, somitogenesis and eye formation. In the adult, Dkks are implicated in bone formation and bone disease, cancer and

		Alzheimer disease (By similarity).
Chr5:1968398-1971603	-	-
Chr5:24370524-24373453	-	-
Chr5:114756499-114757907	-	-
Chr5:143512864-143515050	YIPF5	YIPF5 (Yip1 Domain Family Member 5): Diseases associated with YIPF5 include Pleomorphic Adenoma Carcinoma.
Chr5:176387572-176390195	UIMC1	UIMC1 (Ubiquitin Interaction Motif Containing 1): Among its related pathways are Cell Cycle, Mitotic and Metabolism of proteins. Plays a central role in the BRCA1-A complex by specifically binding Lys-63-linked ubiquitinated histones H2A and H2AX at DNA lesions sites, leading to target the BRCA1-BARD1 heterodimer to sites of DNA damage at double-strand breaks (DSBs). May indirectly act as a transcriptional repressor by inhibiting the interaction of NR6A1 with the corepressor NCOR1.
Chr6:31211614-31213131	-	-
Chr6:56758337-56760961	DST	DST (Dystonin): Diseases associated with DST include Neuropathy, Hereditary Sensory And Autonomic, Type Vi and Epidermolysis Bullosa Simplex [SKIN DISESASE], Autosomal Recessive 2. Among its related pathways are Collagen chain trimerization and Degradation of the extracellular matrix. Required for anchoring either intermediate filaments to the actin cytoskeleton in neural and muscle cells or keratin-containing intermediate filaments to hemidesmosomes in epithelial cells. The proteins may self-aggregate to form filaments or a two-dimensional mesh. Regulates the organization and stability of the microtubule network of sensory neurons to allow axonal transport. Mediates docking of the dynein/dynactin motor complex to vesicle cargos for retrograde axonal transport through its interaction with TMEM108 and DCTN1 (By similarity). Isoform 3: plays a structural role in the assembly of hemidesmosomes of epithelial cells; anchors keratin-containing intermediate filaments to the inner plaque of hemidesmosomes [IN SKIN CELLS]. Isoform 6: required for bundling actin filaments around the nucleus. Isoform 7: regulates the organization and stability of the microtubule network of sensory neurons to allow axonal transport.
Chr6:85318139-85324237	-	-
Chr7:55286741-55289075	(EGFR 50519 bp away)	CLOSE TO EGFR EGFR (Epidermal Growth Factor Receptor): Diseases associated with EGFR include Inflammatory Skin And Bowel Disease, Neonatal, 2 and Lung Cancer. The epidermal growth factor receptor (EGFR) is a receptor tyrosine kinase of the ErbB family. Four members of the ErbB family have been identified; EGFR (ErbB1, HER1), ErbB2 (HER2), ErbB3 (HER3) and ErbB4 (HER4). EGFR signaling drives many cellular responses. Plays a role in enhancing learning and memory performance (By similarity). Isoform 2 may act as an antagonist of EGF action. (Microbial infection) Acts as a receptor for hepatitis C virus (HCV) in hepatocytes and facilitates its cell entry. Mediates HCV

		entry by promoting the formation of the CD81-CLDN1 receptor complexes that are essential for HCV entry and by enhancing membrane fusion of cells expressing HCV envelope glycoproteins.
Chr7:113416156-113422223	-	-
Chr8:18454605-18455959	PSD3	PSD3 (Pleckstrin And Sec7 Domain Containing 3): Among its related pathways are Endocytosis.
Chr8:73787759-73793833	KCNB2	KCNB2 (Potassium Voltage-Gated Channel Subfamily B Member 2): Diseases associated with KCNB2 include Brugada Syndrome [heart disease]. Among its related pathways are Dopamine-DARPP32 Feedback onto cAMP Pathway and Potassium Channels. Voltage-gated potassium channel that mediates transmembrane potassium transport in excitable membranes, primarily in the brain and smooth muscle cells.
Chr8:126595113-126601150	-	-
Chr8:129465149-129471278	-	-
Chr8:135082913-135089028	-	-
Chr9:84324350-84326939	TLE1	TLE1 (Transducin Like Enhancer Of Split 1): Diseases associated with TLE1 include Synovium Cancer and Spindle Cell Liposarcoma [Liposarcoma is a rare cancer of connective tissues that resemble fat cells under a microscope.]. Among its related pathways are Signaling by Wnt and Wnt / Hedgehog / Notch.
Chr9:90858982-90860974	-	-
Chr9:110033433-110035453	-	-
Chr9:110537518-110540610	-	-
Chr10:111572110-111578229	XPNPEP1	XPNPEP1 (X-Prolyl Aminopeptidase 1): Diseases associated with XPNPEP1 include Mature Cataract and Biliary Atresia. Gene Ontology (GO) annotations related to this gene include protein homodimerization activity and manganese ion binding. Contributes to the degradation of bradykinin [Bradykinin is an inflammatory mediator. It is a peptide that causes blood vessels to dilate (enlarge), and therefore causes blood pressure to fall].
Chr12:66527627-66529900	TMBIM4,LLPH	TMBIM4 (Transmembrane BAX Inhibitor Motif Containing 4): is a Protein Coding gene. Diseases associated with TMBIM4 include Venezuelan Hemorrhagic Fever. LLPH (LLP Homolog, Long-Term Synaptic Facilitation Factor): In hippocampal neurons, regulates dendritic and spine growth and synaptic transmission.
Chr12:96233574-96236340	NTN4, SNRPF	NTN4 (Netrin 4): Among its related pathways are Developmental Biology and Non-integrin membrane-ECM interactions. May play an important role in neural, kidney and vascular development. Promotes neurite elongation from olfactory bulb explants. SNRPF (Small Nuclear Ribonucleoprotein Polypeptide F): Among its related pathways are mRNA Splicing - Major Pathway and Processing of Capped Intronless Pre-mRNA. As

		part of the U7 snRNP it is involved in histone 3-end processing.
Chr12:96340340-96342967	AMDHD1, CCDC38, HAL	AMDHD1 (Amidohydrolase Domain Containing 1): Among its related pathways are histidine degradation and Metabolism. CCDC38 (Coiled-Coil Domain Containing 38). HAL (Histidine Ammonia-Lyase): Diseases associated with HAL include Histidinemia and Histidine Metabolism Disease. Among its related pathways are histidine degradation and Metabolism. Histidine ammonia-lyase is a cytosolic enzyme catalyzing the first reaction in histidine catabolism, the nonoxidative deamination of L-histidine to trans-urocanic acid. Histidine ammonia-lyase defects cause histidinemia which is characterized by increased histidine and histamine and decreased urocanic acid in body fluids.
Chr16:46401931-46406443	-	-
Chr18:41031536-41032542	-	-
Chr19:40613089-40615774	ZNF780A	ZNF780A (Zinc Finger Protein 780A).
Chr19:52175921-52177533	SIGLEC14, HAS1	SIGLEC14 (Sialic Acid Binding Ig Like Lectin 14): Among its related pathways are Innate Immune System and RET signaling. HAS1: (Hyaluronan Synthase 1). Diseases associated with HAS1 include Waldenstrom Macroglobulinemia [a type of cancer affecting two types of B cells, lymphoplasmacytoid cells and plasma cells.] and Foodborne Botulism [Foodborne botulism is a food poisoning caused by a toxin produced by the bacterium, Clostridium botulinum.]. Among its related pathways are Metabolism and Glycosaminoglycan metabolism. [I]t is essential to hyaluronan synthesis a major component of most extracellular matrices that has a structural role in tissues architectures and regulates cell adhesion, migration and differentiation.
Chr20:2803143-2806429	TMEM239, PCED1A	TMEM239 (Transmembrane Protein 239). PCED1A (PC-Esterase Domain Containing 1A): [B]elongs to the Pmr5-Cas1p-esterase subfamily in that it contains the catalytic triad comprised of serine, aspartate and histidine and lacks two conserved regions (glycine after strand S2 and GxND motif).
Chr22:27168065-27169805	-	-

Table S5.4: GO-terms with FDR<10% among 35 genes within 50 kb of fixed deletions.

Term	Count	%	PValue	Fold Enrichment	FDR
GO:0019557~histidine catabolic process to glutamate and formate	2 (AMDHD1, HAL)	5.88	0.00689	279.87	8.20
GO:0019556~histidine catabolic process to glutamate and formamide	2 (AMDHD1, HAL)	5.88	0.00689	279.87	8.20
GO:0005604~basement membrane	3 (FRAS1, NTN4, DST)	8.82	0.00795	21.63	8.23

Table S5.5: (A) Variant effect counts and (B) allele frequency estimates in the Khoe-San populations.

(A)

VARIANT EFFECT (COUNT)	IMPACT	Total	Karretjie	Nama	|Gui and || Gana	Ju|'hoansi	!Xun
missense	MODERATE	62,364	30,857	31,143	31,000	30,883	31,298
synonymous	LOW	59,398	32,420	32,005	32,777	31,844	32,911
start-loss	HIGH	86	38	46	39	40	42
stop-gain	HIGH	994	396	346	334	372	327
stop-loss	HIGH	70	41	37	42	37	40
splice acceptor	HIGH	666	333	361	326	330	325
splice donor	HIGH	758	405	402	414	397	397
intergenic	MODIFIER	8,733,644	5,002,169	4,991,530	5,049,016	4,938,062	5,058,988
intronic	MODIFIER	8,176,893	4,581,736	4,559,468	4,621,642	4,534,834	4,638,010

(B)

VARIANT EFFECT (COUNT)	IMPACT	Total	Allele Freq. Mean	Allele Freq. SD
missense	MODERATE	62,364	0.15	0.24
synonymous	LOW	59,398	0.18	0.25
start-loss	HIGH	86	0.15	0.24
stop-gain	HIGH	994	0.07	0.13
stop-loss	HIGH	70	0.19	0.26
splice acceptor	HIGH	666	0.13	0.20
splice donor	HIGH	758	0.16	0.23
intergenic	MODIFIER	8,733,644	NA	NA
intronic	MODIFIER	8,176,893	NA	NA

Table S5.6: Variant effect counts in Khoe-San individuals, including means and standard deviations per population.

Pop.	Sample	Missense	Synonym-ous	Start-loss	Stop-gain	Stop-loss	Splice acceptor	Splice donor	Intergenic	Intronic
GUG	KSP092	14,310	15,547	17	124	17	137	185	2,425,736	2,195,002
GUG	KSP096	14,116	15,392	17	134	19	139	189	2,409,295	2,171,043
GUG	KSP224	13,820	15,578	20	104	18	138	186	2,412,870	2,185,898
GUG	KSP225	14,186	15,698	20	126	23	148	179	2,434,291	2,198,351
GUG	KSP228	14,081	15,433	13	116	19	135	200	2,406,275	2,194,802
JUH	KSP103	14,075	15,138	18	163	15	150	186	2,448,680	2,213,249
JUH	KSP105	14,425	15,731	20	125	22	148	196	2,431,390	2,215,841
JUH	KSP106	14,431	15,471	23	134	21	145	177	2,440,963	2,208,014
JUH	KSP111	14,123	15,445	24	102	15	123	179	2,444,572	2,213,964
JUH	KSP116	14,365	15,689	15	123	19	151	172	2,428,824	2,201,216
XUN	KSP146	14,231	15,622	17	106	20	127	196	2,431,753	2,193,740
XUN	KSP150	14,271	15,451	19	120	23	132	187	2,427,538	2,192,210
XUN	KSP152	14,164	15,639	19	106	18	140	175	2,430,081	2,201,452
XUN	KSP154	14,074	15,494	24	105	22	137	189	2,429,547	2,194,690
XUN	KSP155	14,341	15,685	19	114	18	152	192	2,422,595	2,196,452
KAR	KSP062	13,925	15,426	16	118	24	135	200	2,406,566	2,183,824
KAR	KSP063	14,701	15,826	19	178	21	142	202	2,469,034	2,240,651
KAR	KSP065	14,191	15,521	18	133	21	133	194	2,426,457	2,177,217
KAR	KSP067	13,946	15,459	15	105	17	140	170	2,415,239	2,183,029
KAR	KSP069	14,181	15,618	21	115	19	131	180	2,437,770	2,199,214
NAM	KSP124	14,344	15,313	18	146	18	171	226	2,444,745	2,201,520
NAM	KSP134	14,237	15,513	20	114	17	148	204	2,388,936	2,169,668
NAM	KSP137	14,014	15,081	22	126	20	143	168	2,387,076	2,151,821
NAM	KSP139	13,832	15,139	19	106	18	128	177	2,397,056	2,151,709
NAM	KSP140	14,118	15,319	28	95	20	136	176	2,376,812	2,145,063
KSP MEAN (SD)		14,180.1 (199.4)	15,489.1 (189.0)	19.2 (3.3)	121.5 (19.2)	19.4 (2.4)	140.4 (10.1)	187.4 (13.1)	2,422,964.0 (21269.1)	2,191,185.6 (21,694.1)
GUG MEAN (SD)		14,102.6 (180.6)	15,529.6 (121.7)	17.4 (2.9)	120.8 (11.4)	19.2 (2.3)	139.4 (5)	187.8 (7.7)	2,417,693.4 (11,877.9)	2,189,019.2 (11,059.0)

JUH MEAN (SD)	14,283.8 (171.5)	15,494.8 (236.6)	20.0 (3.7)	129.4 (22.1)	18.4 (3.3)	143.4 (11.6)	182.0 (9.3)	2,438,885.8 (8,514.7)	2,210,456.8 (5,925.5)
XUN MEAN (SD)	14,216.2 (102.2)	15,578.2 (100.4)	19.6 (2.6)	110.2 (6.6)	20.2 (2.3)	137.6 (9.4)	187.8 (7.9)	2,428,302.8 (3,527.7)	2,195,708.8 (3,559.9)
KAR MEAN (SD)	14,188.8 (312.6)	15,570.0 (160.7)	17.8 (2.4)	129.8 (28.8)	20.4 (2.6)	136.2 (4.7)	189.2 (13.8)	2,431,013.2 (24,282.1)	2,196,787.0 (25,837.2)
NAM MEAN (SD)	14,109.0 (198.4)	15,273.0 (170.4)	21.4 (4.0)	117.4 (19.6)	18.6 (1.3)	145.2 (16.3)	190.2 (24.2)	2,398,925.0 (26,608.9)	2,163,956.2 (22,901.3)

Table S5.7: Results of gene set enrichment analysis in the KSP dataset, based on high impact variants. FDR p-val threshold < 0.05.

start_lost_variants **87**

GO term	Ave # genes found / sim	Genes found for GO category	P-value (uncorrected)	P-value (FDR adjusted)	Unique Genes found for GO category	Max # genes for GO category	Total genes for GO category in GO file	Description GO term
GO:0050911	0.101	8	0.00001	0.006516	8	363	499	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	0.153	8	0.00001	0.006516	8	398	554	detection of chemical stimulus involved in sensory perception
GO:0050906	0.675	8	0.00001	0.006516	8	444	603	detection of stimulus involved in sensory perception
GO:0009593	0.424	8	0.00001	0.006516	8	436	597	detection of chemical stimulus
GO:0004930	1.726	11	0.00001	0.006516	11	729	919	G-protein coupled receptor activity
GO:0021984	0.01	2	0.00002	0.01108	2	6	7	adenohypophysis development

stop_gained_variants **1003**

GO term	Ave # genes found / sim	Genes found for GO category	P-value (uncorrected)	P-value (FDR adjusted)	Unique Genes found for GO category	Max # genes for GO category	Total genes for GO category in GO file	Description GO term
GO:0005882	3.013	19	0.00001	0.0027942857	19	182	230	intermediate filament
GO:0001580	0.405	5	0.00001	0.0027942857	5	28	45	detection of chemical stimulus involved in sensory perception of bitter taste
GO:0045095	0.552	9	0.00001	0.0027942857	9	91	121	keratin filament
GO:0038023	55.765	91	0.00001	0.0027942857	91	1162	1518	signaling receptor activity
GO:0050912	0.531	6	0.00001	0.0027942857	6	34	54	detection of chemical stimulus involved in sensory perception of taste
GO:0050911	1.145	52	0.00001	0.0027942857	52	363	499	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	1.722	58	0.00001	0.0027942857	58	398	554	detection of chemical stimulus involved in sensory perception

GO:0050906	6.501	61	0.00001	0.0027942857	61	444	603	detection of stimulus involved in sensory perception
GO:0009593	4.494	59	0.00001	0.0027942857	59	436	597	detection of chemical stimulus
GO:0004984	0.409	23	0.00001	0.0027942857	23	151	190	olfactory receptor activity
GO:0004930	18.204	66	0.00001	0.0027942857	66	729	919	G-protein coupled receptor activity
GO:0004888	47.514	84	0.00001	0.0027942857	84	1064	1410	transmembrane signaling receptor activity
GO:0051606	18.753	75	0.00001	0.0027942857	75	647	833	detection of stimulus
GO:0004872	65.3	102	0.00001	0.0027942857	102	1344	1750	receptor activity
GO:0016712	0.781	6	0.00007	0.0150617647	6	25	28	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen
GO:0009404	0.277	4	0.00007	0.0150617647	4	11	11	toxin metabolic process
GO:0042611	0.789	6	0.00007	0.0150617647	6	20	123	MHC protein complex
GO:0046977	0.121	3	0.00009	0.0203410526	3	5	39	TAP binding
GO:0030156	0.026	2	0.00009	0.0203410526	2	2	2	benzodiazepine receptor binding
GO:0060335	0.164	3	0.00013	0.0251752381	3	4	4	positive regulation of interferon-gamma-mediated signaling pathway
GO:0060332	0.164	3	0.00013	0.0251752381	3	4	4	positive regulation of response to interferon-gamma
GO:0001909	0.585	5	0.00015	0.02817	5	25	39	leukocyte mediated cytotoxicity

stop_lost_variants **70**

GO term	Ave # genes found / sim	Genes found for GO category	P-value (uncorrected)	P-value (FDR adjusted)	Unique Genes found for GO category	Max # genes for GO category	Total genes for GO category in GO file	Description GO term
GO:0050911	0.089	5	0.00001	0.01576	5	363	499	detection of chemical stimulus involved in

								sensory perception of smell
GO:0050907	0.132	5	0.00001	0.01576	5	398	554	detection of chemical stimulus involved in sensory perception

splice_acceptor_variants	**680**
None significant	

splice_donor_variants	**777**
None significant	

missense_variants_CADD20 16381

GO term	Ave # genes found / sim	Genes found for GO category	P-value (uncorrected)	P-value (FDR adjusted)	Unique Genes found for GO category	Max # genes for GO category	Total genes for GO category in GO file	Description GO term
GO:0005882	33.743	78	0.00001	0.0012636	78	182	230	intermediate filament
GO:0034660	152.437	190	0.00001	0.0012636	190	380	419	ncRNA metabolic process
GO:0005730	309.993	358	0.00001	0.0012636	358	725	821	nucleolus
GO:0034470	97.014	127	0.00001	0.0012636	127	247	277	ncRNA processing
GO:0005198	224.307	261	0.00001	0.0012636	261	539	605	structural molecule activity
GO:0038023	400.858	576	0.00001	0.0012636	576	1162	1518	signaling receptor activity
GO:0050911	16.877	212	0.00001	0.0012636	212	363	499	detection of chemical stimulus involved in sensory perception of smell
GO:0071025	2.23	7	0.00001	0.0012636	7	7	12	RNA surveillance
GO:0050907	23.916	221	0.00001	0.0012636	221	398	554	detection of chemical stimulus involved in sensory perception
GO:0050906	51.208	248	0.00001	0.0012636	248	444	603	detection of stimulus involved in sensory perception
GO:0004984	6.884	88	0.00001	0.0012636	88	151	190	olfactory receptor activity
GO:0004930	158.445	349	0.00001	0.0012636	349	729	919	G-protein coupled receptor activity
GO:0004888	343.843	532	0.00001	0.0012636	532	1064	1410	transmembrane signaling receptor activity
GO:0004872	491.517	656	0.00001	0.0012636	656	1344	1750	receptor activity

GO:0004871	537.754	685	0.00001	0.0012636	685	1416	1800	signal transducer activity
GO:0045095	10.545	38	0.00001	0.0012636	38	91	121	keratin filament
GO:0007608	16.956	30	0.00001	0.0012636	30	56	60	sensory perception of smell
GO:0007606	28.585	47	0.00001	0.0012636	47	90	96	sensory perception of chemical stimulus
GO:0007519	20.458	33	0.00001	0.0012636	33	49	51	skeletal muscle tissue development
GO:0006364	29.726	51	0.00001	0.0012636	51	96	105	rRNA processing
GO:0009593	44.563	240	0.00001	0.0012636	240	436	597	detection of chemical stimulus
GO:0051606	154.733	338	0.00001	0.0012636	338	647	833	detection of stimulus
GO:0044430	632.06	694	0.00001	0.0012636	694	1271	1412	cytoskeletal part
GO:0060089	626.71	763	0.00001	0.0012636	763	1594	2028	molecular transducer activity
GO:0016072	32.791	55	0.00001	0.0012636	55	102	111	rRNA metabolic process
GO:0021546	0.769	5	0.00002	0.0025142308	5	7	7	rhombomere development
GO:0003774	72.833	89	0.00003	0.0033968966	89	111	120	motor activity
GO:0043228	1088.309	1161	0.00003	0.0033968966	1161	2194	2462	non-membrane-bounded organelle
GO:0043232	1088.309	1161	0.00003	0.0033968966	1161	2194	2462	intracellular non-membrane-bounded organelle
GO:0004980	0.235	3	0.00004	0.0040066667	3	3	3	melanocyte-stimulating hormone receptor activity
GO:0004812	15.4	26	0.00004	0.0040066667	26	35	39	aminoacyl-tRNA ligase activity
GO:0016875	15.4	26	0.00004	0.0040066667	26	35	39	ligase activity, forming carbon-oxygen bonds
GO:0016876	15.4	26	0.00004	0.0040066667	26	35	39	ligase activity, forming aminoacyl-tRNA and related compounds
GO:0032982	12.892	20	0.00005	0.0049373529	20	21	23	myosin filament
GO:0071027	1.684	5	0.00006	0.0055102703	5	5	10	nuclear RNA surveillance
GO:0071028	1.684	5	0.00006	0.0055102703	5	5	10	nuclear mRNA surveillance
GO:0032391	15.755	23	0.00007	0.0055102703	23	29	31	photoreceptor connecting cilium
GO:0004386	55.014	71	0.00008	0.0070153846	71	99	126	helicase activity
GO:0042255	4.078	10	0.00008	0.0070153	10	12	14	ribosome assembly

				846				
GO:0006418	21.694	33	0.00009	0.0075414634	33	46	51	tRNA aminoacylation for protein translation
GO:0021570	0.159	3	0.00009	0.0075414634	3	3	3	rhombomere 4 development
GO:0045111	23.47	34	0.0001	0.008295	34	50	59	intermediate filament cytoskeleton
GO:0034663	3.962	9	0.00011	0.00865	9	11	11	endoplasmic reticulum chaperone complex
GO:0043038	23.278	35	0.00011	0.00865	35	49	54	amino acid activation
GO:0043039	23.278	35	0.00011	0.00865	35	49	54	tRNA aminoacylation
GO:0016459	36.643	47	0.00013	0.0093028261	47	61	65	myosin complex
GO:0005814	50.427	64	0.00016	0.0123744681	64	90	99	centriole
GO:0008188	15.961	24	0.00019	0.0145508333	24	38	43	neuropeptide receptor activity
GO:0070391	2.595	7	0.0002	0.0147158	7	9	9	response to lipoteichoic acid
GO:0071223	2.595	7	0.0002	0.0147158	7	9	9	cellular response to lipoteichoic acid
GO:0016887	135.595	157	0.00026	0.0182615686	157	241	297	ATPase activity
GO:0051181	9.005	16	0.00029	0.0195351786	16	21	22	cofactor transport
GO:0044450	69.129	84	0.00029	0.0195351786	84	123	138	microtubule organizing center part
GO:0014823	21.086	31	0.0003	0.0195351786	31	54	63	response to activity
GO:0008544	28.175	40	0.0003	0.0195351786	40	81	90	epidermis development
GO:0045684	11.922	19	0.0003	0.0195351786	19	31	37	positive regulation of epidermis development
GO:0005859	6.609	13	0.00033	0.021752069	13	17	18	muscle myosin complex
GO:0015926	5.131	9	0.00033	0.021752069	9	9	10	glucosidase activity
GO:2000543	2.169	6	0.00034	0.021970678	6	7	10	positive regulation of gastrulation
GO:0009952	51.945	66	0.0004	0.0259638333	66	135	149	anterior/posterior pattern specification
GO:0042633	2.293	8	0.00042	0.0261295455	8	15	16	hair cycle
GO:0042303	2.293	8	0.00042	0.0261295455	8	15	16	molting cycle
GO:0050840	19.037	28	0.00044	0.0261295455	28	43	49	extracellular matrix binding

GO:0070035	22.891	33	0.00044	0.0261295455	33	46	63	purine NTP-dependent helicase activity
GO:0016634	4.51	9	0.00044	0.0261295455	9	9	9	oxidoreductase activity, acting on the CH-CH group of donors, oxygen as acceptor
GO:0008026	22.891	33	0.00044	0.0261295455	33	46	63	ATP-dependent helicase activity
GO:0030968	36.263	50	0.00045	0.0262994203	50	98	104	endoplasmic reticulum unfolded protein response
GO:0006399	71.213	90	0.00046	0.0262994203	90	167	192	tRNA metabolic process
GO:0014707	0.265	3	0.00046	0.0262994203	3	3	3	branchiomeric skeletal muscle development
GO:0008186	7.659	15	0.00047	0.0265015714	15	21	33	RNA-dependent ATPase activity
GO:0015884	2.019	6	0.00049	0.0268647222	6	7	7	folic acid transport
GO:0036498	21.597	32	0.00049	0.0268647222	32	55	56	IRE1-mediated unfolded protein response
GO:0008537	0.359	3	0.0005	0.0269354667	3	3	3	proteasome activator complex
GO:0016801	3.294	8	0.00051	0.0269354667	8	9	9	hydrolase activity, acting on ether bonds
GO:0021568	0.05	2	0.00052	0.0269354667	2	2	2	rhombomere 2 development
GO:0046128	81.583	100	0.00056	0.0286790789	100	188	206	purine ribonucleoside metabolic process
GO:0044262	59.732	75	0.00058	0.0297896104	75	131	137	cellular carbohydrate metabolic process
GO:0000076	4.958	10	0.00059	0.0299008974	10	11	12	DNA replication checkpoint
GO:0042623	97.621	115	0.00062	0.0310837975	115	173	220	ATPase activity, coupled
GO:0006268	1.761	6	0.00067	0.0329898765	6	8	8	DNA unwinding involved in DNA replication
GO:0007413	11.812	16	0.00067	0.0329898765	16	19	19	axonal fasciculation
GO:0006805	77.962	97	0.00072	0.0353212195	97	179	192	xenobiotic metabolic process
GO:0042923	1.954	5	0.00074	0.0359308434	5	6	6	neuropeptide binding
GO:0044453	3.135	7	0.00077	0.0370560714	7	7	7	nuclear membrane part
GO:0016460	9.757	16	0.0008	0.0375226667	16	21	22	myosin II complex
GO:0009888	243.038	270	0.00081	0.0375226	270	533	571	tissue development

				667				
GO:0032508	25.463	35	0.00083	0.0375226667	35	48	55	DNA duplex unwinding
GO:0000226	141.008	160	0.00083	0.0375226667	160	243	268	microtubule cytoskeleton organization
GO:0046364	18.734	28	0.00083	0.0375226667	28	51	56	monosaccharide biosynthetic process
GO:0045494	23.302	31	0.00083	0.0375226667	31	38	40	photoreceptor cell maintenance
GO:0032527	4.558	10	0.00086	0.0384943956	10	15	15	protein exit from endoplasmic reticulum
GO:0009126	52.588	67	0.0009	0.0399218478	67	125	140	purine nucleoside monophosphate metabolic process
GO:0000178	6.082	12	0.00096	0.0421755914	12	17	17	exosome (RNase complex)
GO:0021912	0.099	2	0.001	0.0438094681	2	2	2	regulation of transcription from RNA polymerase II promoter involved in spinal cord motor neuron fate specification
GO:0005200	41.336	53	0.00109	0.0474765263	53	93	108	structural constituent of cytoskeleton
GO:0032963	46.658	57	0.00111	0.0477891667	57	80	94	collagen metabolic process
GO:0021754	0.067	2	0.00117	0.0498897	2	2	2	facial nucleus development
GO:0005874	174.998	196	0.00118	0.0498897	196	318	358	microtubule
GO:0002583	0.686	4	0.00118	0.0498897	4	6	17	regulation of antigen processing and presentation of peptide antigen
GO:0005875	58.087	70	0.00119	0.0498897	70	100	108	microtubule associated complex

Table S5.8: List of common start-loss variants in the Khoe-San dataset (allele frequency ≥ 10%).

Chr	Position	RS Number	Ref. Allele	Alt. Allele	Allele Freq.	Gene	Global Minor Allele	Global Minor Allele Freq.
1	12853378	rs112330886	T	G	0.1	*PRAMEF1*	G	0.0010
1	153513900	rs7529714	T	C	0.1	*S100A5*	C	0.0254
2	71163086	rs11681642	T	C	0.42	*ATP6V1B1*	C	0.3688
2	99226173	rs1062847	T	A	0.72	*UNC50*	A	0.2857
3	46449164	rs11574440	G	A	0.14	*CCRL2*	A	0.0555
3	100712249	rs3732895	T	C	0.26	*ABI3BP*	C	0.2394
3	190167596	rs543701166	C	T	0.1	*TMEM207*	T	
4	68829109	rs977728	C	T	0.14	*TMPRSS11A*	T	0.2033
4	100485255	rs11944752	G	A	0.48	*MTTP*	A	0.2500
7	1878377	rs3889573	A	G	0.82	*AC110781.3*	A	0.1635
7	30915263	rs10216063	A	G	0.88	*AQP1*	G	0.4597
7	37956138	rs113389918	A	G	0.14	*SFRP4*	G	0.0001
9	21368008	rs146195386	A	G	0.14	*IFNA13*	G	0.0022
9	98638288	rs690528	A	G	0.1	*ERCC6L2*	G	0.2524
11	56113516	rs1905055	T	C	0.92	*OR8K1*	T	0.2200
11	57798427	rs113094669	G	A	0.32	*OR6Q1*	A	0.0002
11	64937708	rs141151195	T	C	0.14	*SPDYC*	C	0.0012
11	77734294	rs78436350	A	T	0.32	*KCTD14*	T	0.0735
14	50472515	rs113920014	C	T	0.34	*C14orf182*	T	0.0078
14	105196232	rs150558753	G	C	0.18	*ADSSL1*	C	0.0027
15	77176158	rs3812908	T	C	0.26	*SCAPER*	C	0.4177
15	81201994	rs28450224	T	C	0.1	*RP11-351M8.1*	C	0.0793
17	28268817	rs79556405	G	A	0.14	*EFCAB5*	A	0.0665
17	68093044	rs7224070	A	G	0.56	*KCNJ16*	A	0.1689
19	2255311	rs7250822	C	G	0.3	*JSRP1*	G	0.2348
19	52223180	rs56166910	A	G	0.4	*HAS1*	G	0.4417
19	54974774	rs1981829	A	T	1	*LENG9*	A	0.4064
19	58144715	rs9749449	A	G	0.52	*ZNF211*	G	0.2145
20	31756954	rs17124277	G	A	0.36	*BPIFA2*	A	0.0280
22	23114327	rs6003299	T	C	0.86	*IGLV3-12*	T	0.3914

Table S5.9: List of common stop-loss variants in the Khoe-San dataset (allele frequency ≥ 10%).

Chr	Position	RS Number	Ref. Allele	Alt. Allele	Allele Freq.	Gene	Global Minor Allele	Global Minor Allele Freq.
1	11906068	rs5065	A	G	0.54	*NPPA*	G	0.1791
2	85549868	rs4240199	A	G	0.88	*TGOLN2*	A	0.2210
2	172180771	rs10205459	A	G	1	*METTL8*	A	0.0018
3	25835983	rs9851096	T	C	0.1	*OXSM*	C	0.0463
3	97806944	rs80220955	T	C	0.14	*OR5AC2*	C	0.1108
4	152212603	rs2407221	T	G	0.6	*PRSS48*	T	0.2204
5	1814783	rs4147773	T	C	0.44	*NDUFS6*	T	0.3870
6	32796685	rs241448	A	G	0.16	*TAP2*	G	0.3111
6	133004281	rs61729583	A	G	0.2	*VNN1*	G	0.0296
7	34889222	rs10275028	T	C	0.36	*NPSR1*	C	0.2672
8	8046160		A	G	0.609	*LRLE1*		
9	116800	rs79220013	C	G	0.22	*FOXD4*	G	0.2031
10	27687225	rs2505323	A	G	0.76	*PTCHD3*	A	0.4499
11	5877979	rs12419602	T	A	0.16	*OR52E8*	A	0.3748
11	55036812	rs61894893	T	C	0.3	*TRIM48*	C	0.0483
11	57983194	rs7103033	A	G	0.14	*OR1S1*	G	0.4335
16	28353929	rs28375936	T	C	0.12	*NPIPB6*		
17	20768730	rs4605228	A	G	0.14	*CCDC144NL*	G	0.0102
18	61379838	rs4940595	T	G	0.52	*SERPINB11*	G	0.3400
19	1783027	rs60482625	T	C	0.34	*ATP8B3*	C	0.0897
19	12540971	rs28559848	T	A	0.22	*ZNF443*	A	0.2768
19	33370070	rs745961	A	G	0.18	*CEP89*	G	0.1643
19	49538868	rs200329830	T	G	0.34	*CGB1*	G	0.0120
19	56499279	rs306457	G	C	0.96	*NLRP8*	G	0.2800
20	60963055	rs911077	T	C	0.16	*RPS21*	C	0.1264
21	45994841	rs7276273	A	C	0.34	*KRTAP10-4*	C	0.0883
22	20710985	rs12160675	A	C	0.36	*FAM230A*	C	0.4491

Table S5.10: List of common stop-gain variants in the Khoe-San dataset (allele frequency ≥ 40%).

Chr	Position	RS Number	Ref. Allele	Alt. Allele	Allele Freq.	Gene	Global Minor Allele	Global Minor Allele Freq.
1	47080679	rs6671527	G	A	0.44	*MOB3C*	G	0.3578
1	158549492	rs863362	C	T	0.5	*OR10X1*	C	0.4976
1	171112490	rs1736565	C	T	0.98	*FMO6P*	C	0.3620
*1	171178090	rs6661174	T	C	0.6	*FMO2*	C	0.0463
2	198593260	rs74375706	A	C	0.56	*BOLL*	C	0.0072
2	228476140	rs2176186	C	T	0.4	*C2orf83*	T	0.3139
2	240323661	rs1709851	C	A	0.76	*AC062017.1*	C	0.2382
3	194061907	rs4974539	G	A	0.46	*CPN2*	A	0.2802
4	70898922	rs17147990	T	A	0.5	*HTN3*	A	0.0473
5	2755485	rs62333235	C	T	0.48	*C5orf38*	T	0.2967
5	134782450	rs12520799	T	A	0.94	*C5orf20*	T	0.3548
6	31124849	rs3130453	C	T	0.64	*CCHCR1*	T	0.4692
6	32552143	rs9269958	C	T	0.833	*HLA-DRB1*	C	0.3057
7	21582963	rs2285943	G	T	0.82	*DNAH11*	T	0.4119
7	64438667	rs1404453	G	A	0.68	*ZNF117*	G	0.1182
7	142231625	rs17249	C	A	0.44	*TRBV10-1*	A	0.2486
8	125579990	rs6470252	C	T	0.86	*NDUFB9*	C	0.3754
9	139937799	rs10781536	G	A	1	*NPDC1*	G	0.0030
11	60265002	rs2298553	C	T	0.6	*MS4A12*	T	0.4782
11	62369881	rs35156678	G	A	0.42	*EML3*	A	0.1655
11	104763117	rs497116	G	A	0.52	*CASP12*	G	0.0515
11	124056732	rs2512227	T	G	0.52	*OR10D3*	G	0.4914
12	40834955	rs10784618	C	A	0.62	*MUC19*	C	0.4816
12	40873278	rs11176811	T	A	0.5	*MUC19*	A	0.1731
12	40873944	rs11176815	C	G	0.5	*MUC19*	G	0.1731
12	55641255	rs4522268	C	T	0.42	*OR6C74*	T	0.2356
14	25103414	rs2273844	G	A	0.64	*GZMB*	A	0.2959
14	106405706	rs11849619	G	A	0.6	*IGHV6-1*	A	0.0915
15	66976378	rs112056079	G	T	0.46	*RP11-321F6.1*	T	0.0166
16	33629700	rs2019670	G	A	0.66	*IGHV3OR16-13*	A	0.3760
16	33629973	rs2002923	G	A	0.64	*IGHV3OR16-13*	A	0.3770
17	39197499	rs3213755	G	A	0.42	*KRTAP1-1*	A	0.1633

19	35719020	rs541169	C	T	0.42	*FAM187B*	T	0.3183
19	52096053	rs3794983	T	A	0.58	*AC018755.1*	A	0.2073
19	57642782	rs9973206	C	A	1	*USP29*	C	0.0479
21	10942756	rs1810540	G	A	0.42	*TPTE*	A	0.3620
22	22707728	rs148013584	C	T	0.42	*IGLV5-48*	T	0.1990

* More recent annotations have come to label rs6661174 as a stop-gain variant, as the C allele is ancestral.

Table S5.11: KSP sample information and mtDNA and Y-chromosome haplogroups.

ID	Population	Language	mtDNA haplogroup	Main mtDNA haplogroup	Y-chromosome haplogroup
KSP062	Karretjie	Indoeuropean*	L0d2a1a	L0d2	E1b1a1a1f1a1d
KSP063	Karretjie	Indoeuropean*	L0d2a1a	L0d2	female
KSP065	Karretjie	Indoeuropean*	L0d2a1a	L0d2	female
KSP067	Karretjie	Indoeuropean*	L0d2a1a	L0d2	E1b1a1a1f1a1*
KSP069	Karretjie	Indoeuropean*	L0d3b1	L0d3	E1b1a1a1g1a2
KSP092	|Gui and ||Gana	Khoe	L0d1c1a2	L0d1	E1b1a1a1g1a1
KSP096	|Gui and ||Gana	Khoe	L0d1c1a1a1	L0d1	E1b1a1a1g1a1
KSP103	Ju|'hoansi	Kx'a	L0d1c3	L0d1	A2a1b1
KSP105	Ju|'hoansi	Kx'a	L0d1c3	L0d1	A3b1c
KSP106	Ju|'hoansi	Kx'a	L0d1a1b	L0d1	A3b1c
KSP111	Ju|'hoansi	Kx'a	L0d1c1a1b	L0d1	B2b1a2
KSP116	Ju|'hoansi	Kx'a	L4b2a2c	L4	A2a1b2
KSP124	Nama	Khoe	L0d2a1a	L0d2	A3b1a
KSP134	Nama	Khoe	L0d2c1a	L0d2	female
KSP137	Nama	Khoe	L0d2a1a	L0d2	E1b1b1d*
KSP139	Nama	Khoe	L0d1a1b1a	L0d1	A3b1a
KSP140	Nama	Khoe	L0d2a1	L0d2	A3b1a
KSP146	!Xun	Kx'a	L0k1a1b	L0k	A3b1b
KSP150	!Xun	Kx'a	L0k1a1a	L0k	A3b1b
KSP152	!Xun	Kx'a	L0d1c3	L0d1	E1b1b1d*
KSP154	!Xun	Kx'a	L0d1c1a1b	L0d1	A2a1b1
KSP155	!Xun	Kx'a	L0k1a1b	L0k	B2b1a2
KSP224	|Gui and ||Gana	Khoe	L0d2c2a	L0d2	female
KSP225	|Gui and ||Gana	Khoe	L0d1c1a1a1	L0d1	female
KSP228	|Gui and ||Gana	Khoe	L1c2b1b	L1c	female

*used to speak a Tuu language

Table S5.12: mtDNA haplogroup frequencies in different Khoe-San speaking populations.

Population	Country	Linguistic affiliation	L0d1	L0d2	L0d3	L0k	L1c	L4	n	Ref
!Xun	Botswana	Kx'a	44,4%	11,1%	0,0%	33,3%	0,0%	0,0%	27	(Barbieri et al. 2014)
G|ui-G||ana	Botswana	Khoe	60,9%	32,6%	0,0%	4,3%	0,0%	0,0%	46	(Barbieri et al. 2014)
Ju|'hoansi	Botswana	Kx'a	50,0%	21,4%	0,0%	23,8%	0,0%	3,6%	84	(Barbieri et al. 2014)
Karretjie	South Africa	Indo-European	25,8%	58,1%	16,1%	0,0%	0,0%	0,0%	31	(Schlebusch et al. 2011)
Nama	Namibia	Khoe	37,9%	34,5%	6,9%	3,4%	0,0%	0,0%	29	(Barbieri et al. 2014)
Taa West	Botswana	Tuu	51,6%	22,6%	0,0%	22,6%	0,0%	0,0%	31	(Barbieri et al. 2014)
Hadza	Tanzania	Khoisan (isolated)	0,0%			0,0%	0,0%	60,0%	79	(Tishkoff et al. 2007)
Sandawe	Tanzania	Khoisan (isolated)	5,0%			0,0%	0,0%	48,0%	82	(Tishkoff et al. 2007)

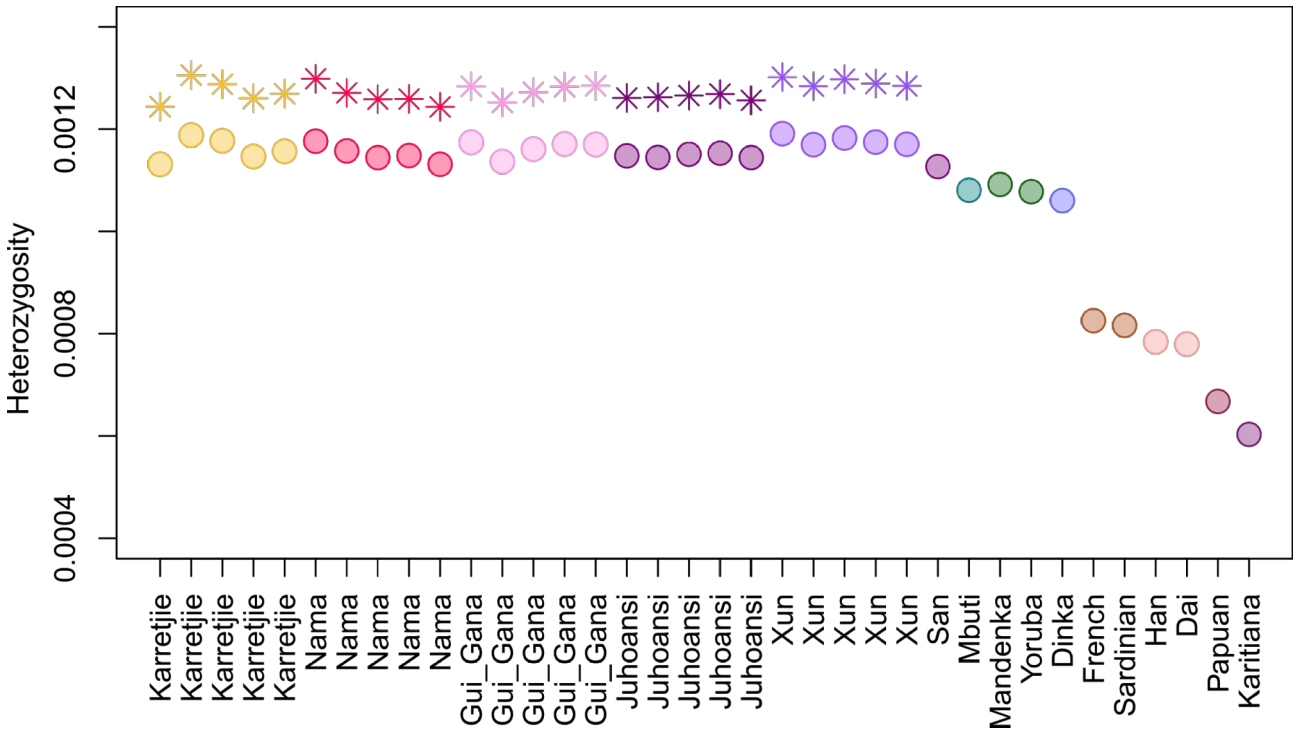


Figure S5.1: Observed heterozygosity in 25 new Khoe-San genomes, compared to five published African and six published non-African genomes (Meyer et al. 2012). Circles indicate estimates done on the KSP+HGDP combined dataset and stars on the Khoe-San only dataset.

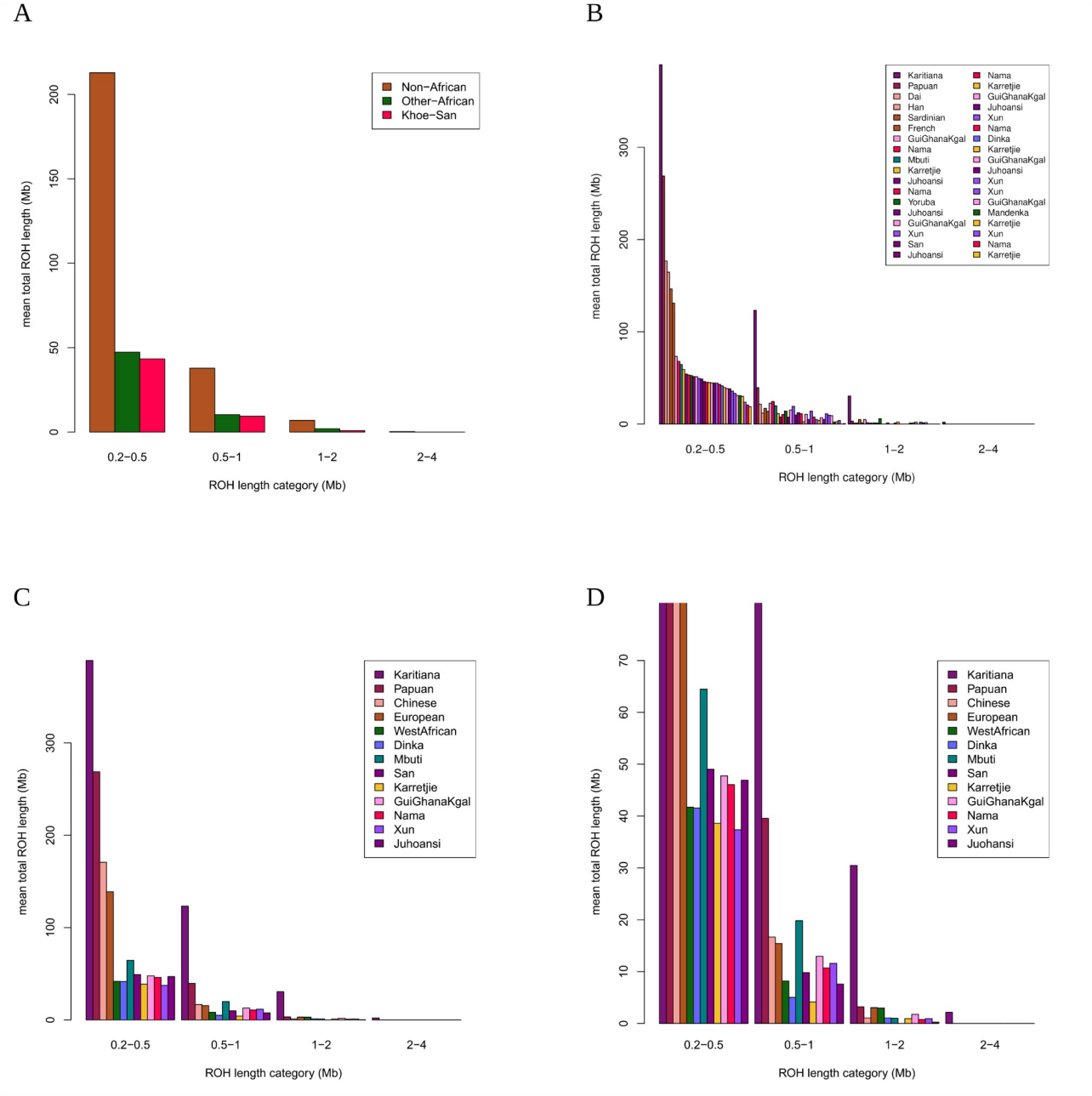


Figure S5.2: Runs of Homozygosity. (A) Khoe-San vs. Other Africans vs. Non-African. (B) Separate individuals in global comparative dataset, sorted descending according to the shortest length class. (order reflected in legend) (C) RoH per group (D) RoH per group zoomed.

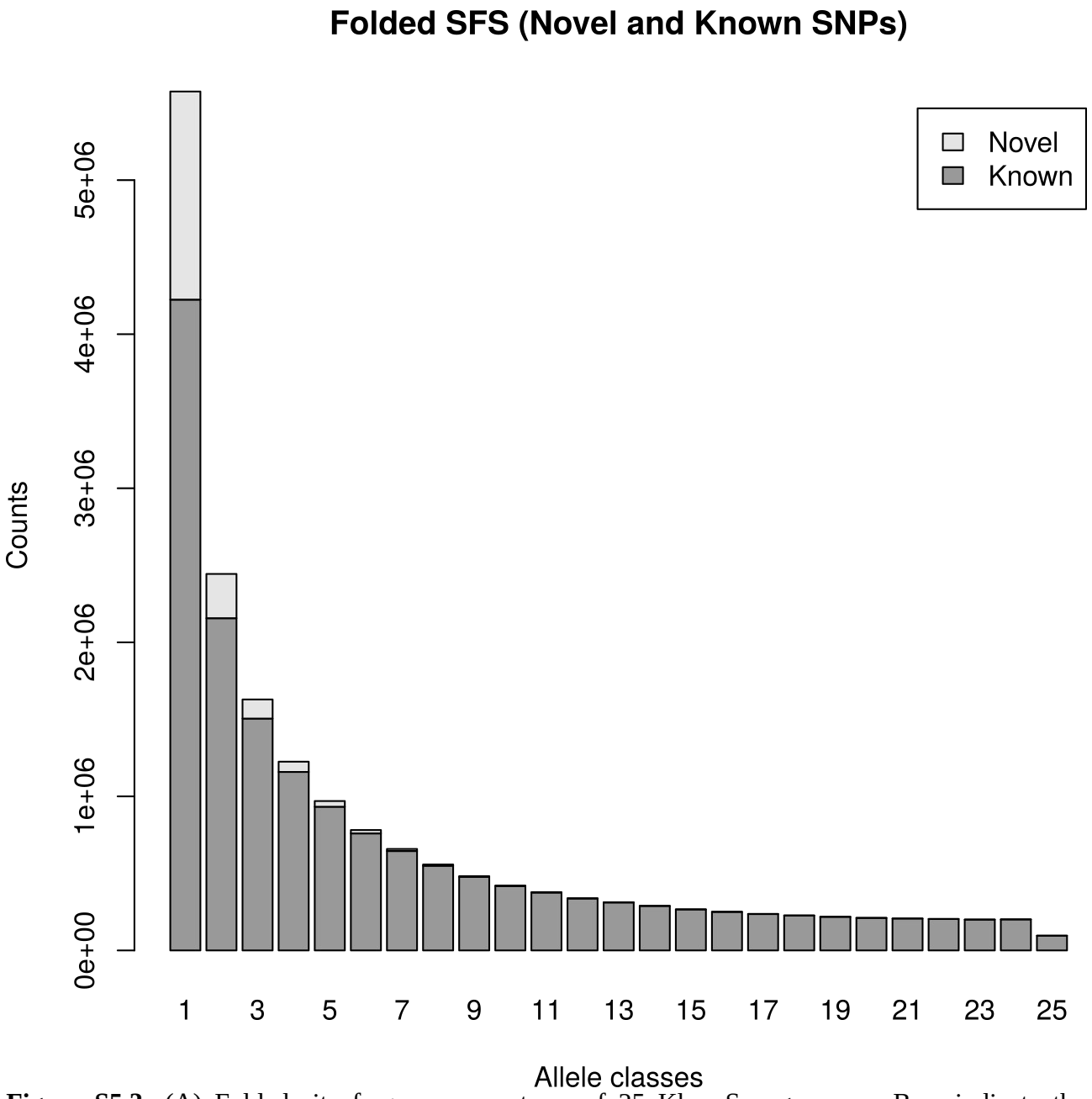


Figure S5.3: (A) Folded site frequency spectrum of 25 Khoe-San genomes. Bars indicate the number of variants in each frequency class. Areas shaded in light gray represent novel variation (compared to dbSNP v.151) and areas shaded in darker gray, known variation.

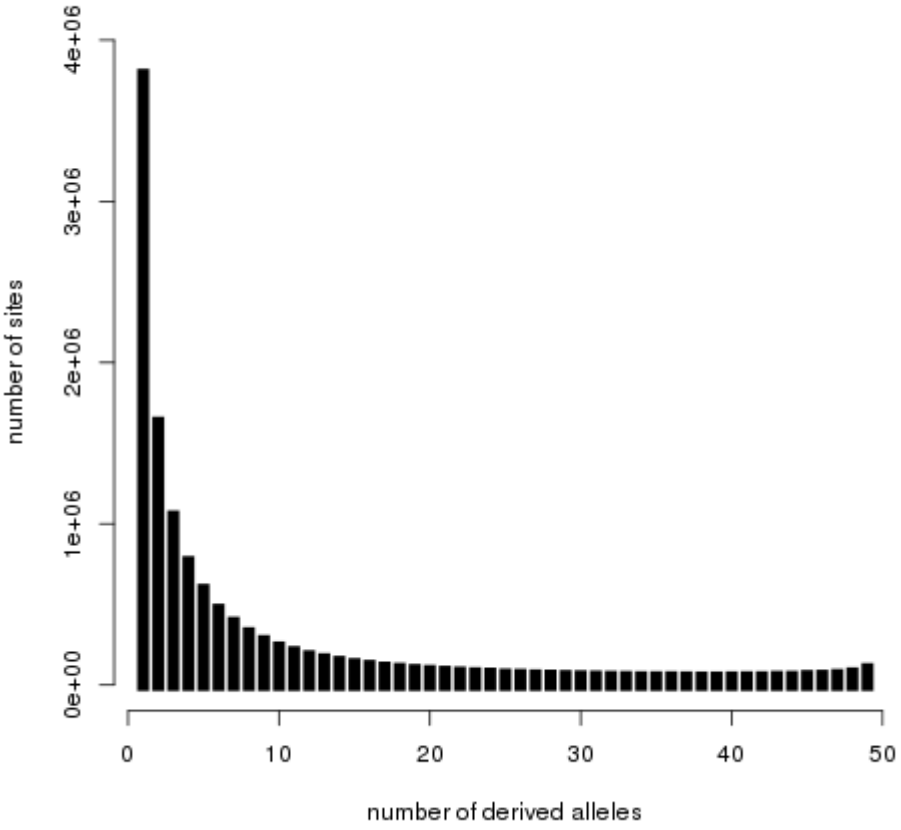


Figure S5.3: (B) Unfolded Site frequency spectrum of 25 Khoe-San genomes.

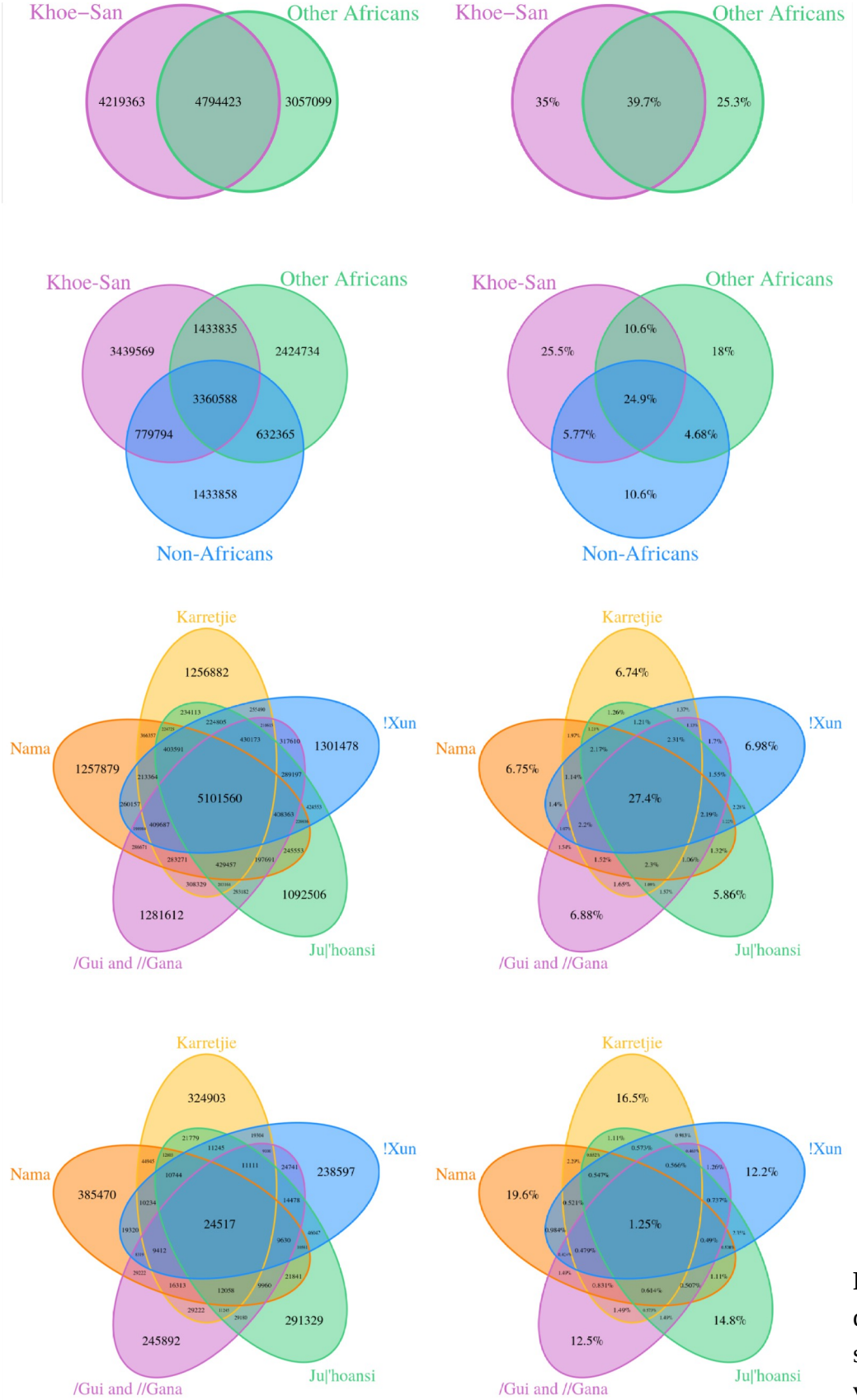


Figure S5.4: Venn diagrams summarizing variants private and shared between different subsets of groups. First column: SNPs counts, second column: percentages of total SNPs count. First row: private and shared variants in Khoe-San vs Other Africans in the HGDP+KSP merged dataset. Second row: private and shared variants in Khoe-San vs Other Africans vs Non-Africans in the HGDP+KSP merged dataset. Third row: private and shared variants in the five Khoe-San populations (KSP dataset). Fourth row: private and shared novel variants in the five

Khoe-San populations (KSP dataset). dbSNP v.151 was used to determine known and novel variants.

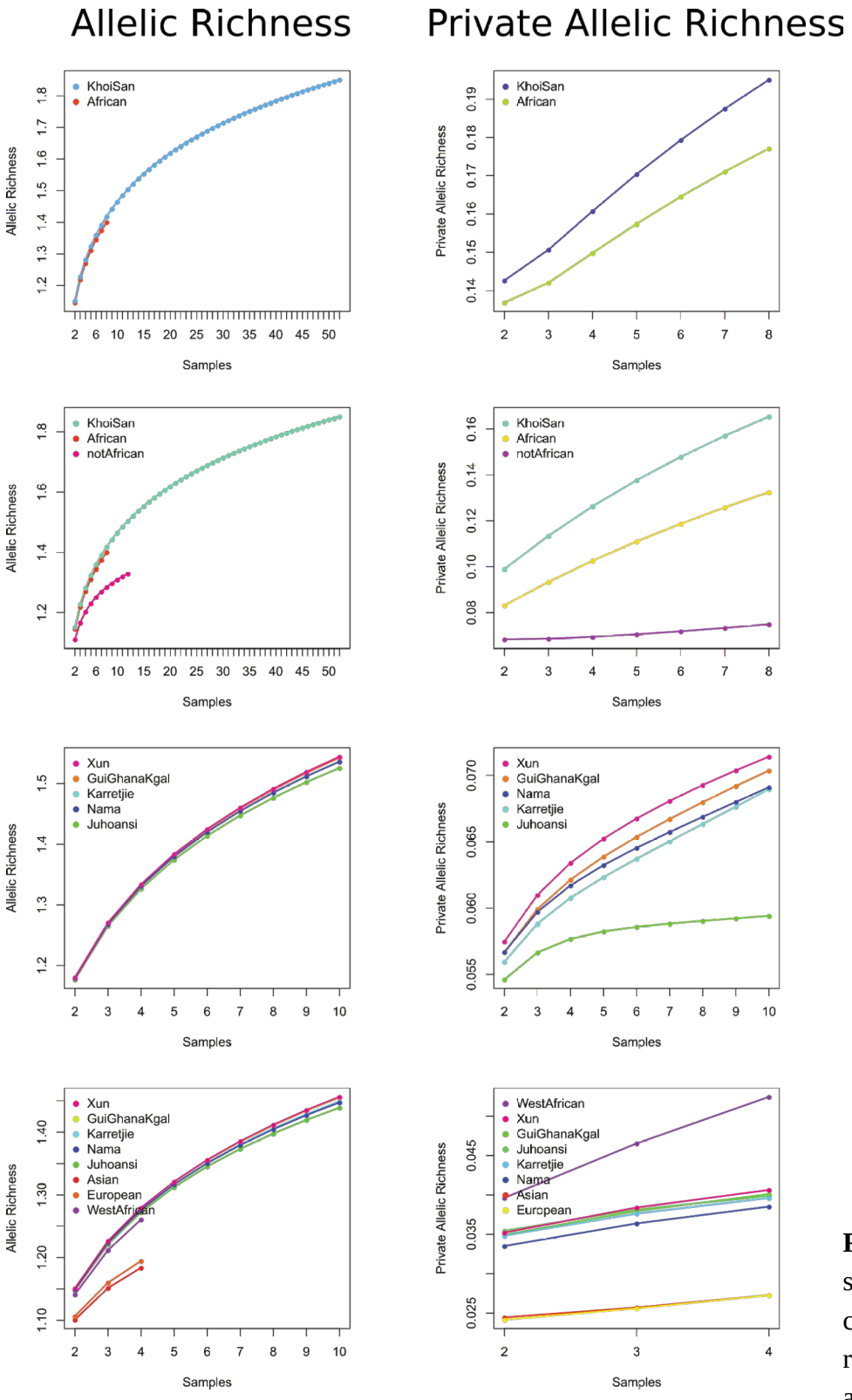


Figure S5.5: Allele sharing (ADZE). Left column: Allelic richness, right column: private allelic richness. First row: Khoe-San vs Other Africans in the HGDP+KSP merged dataset. Second row: Khoe-San vs Other Africans vs Non-Africans in the HGDP+KSP merged dataset. Third row: five Khoe-San populations (KSP dataset). Fourth row: five Khoe-San populations and Asian, European and western African.

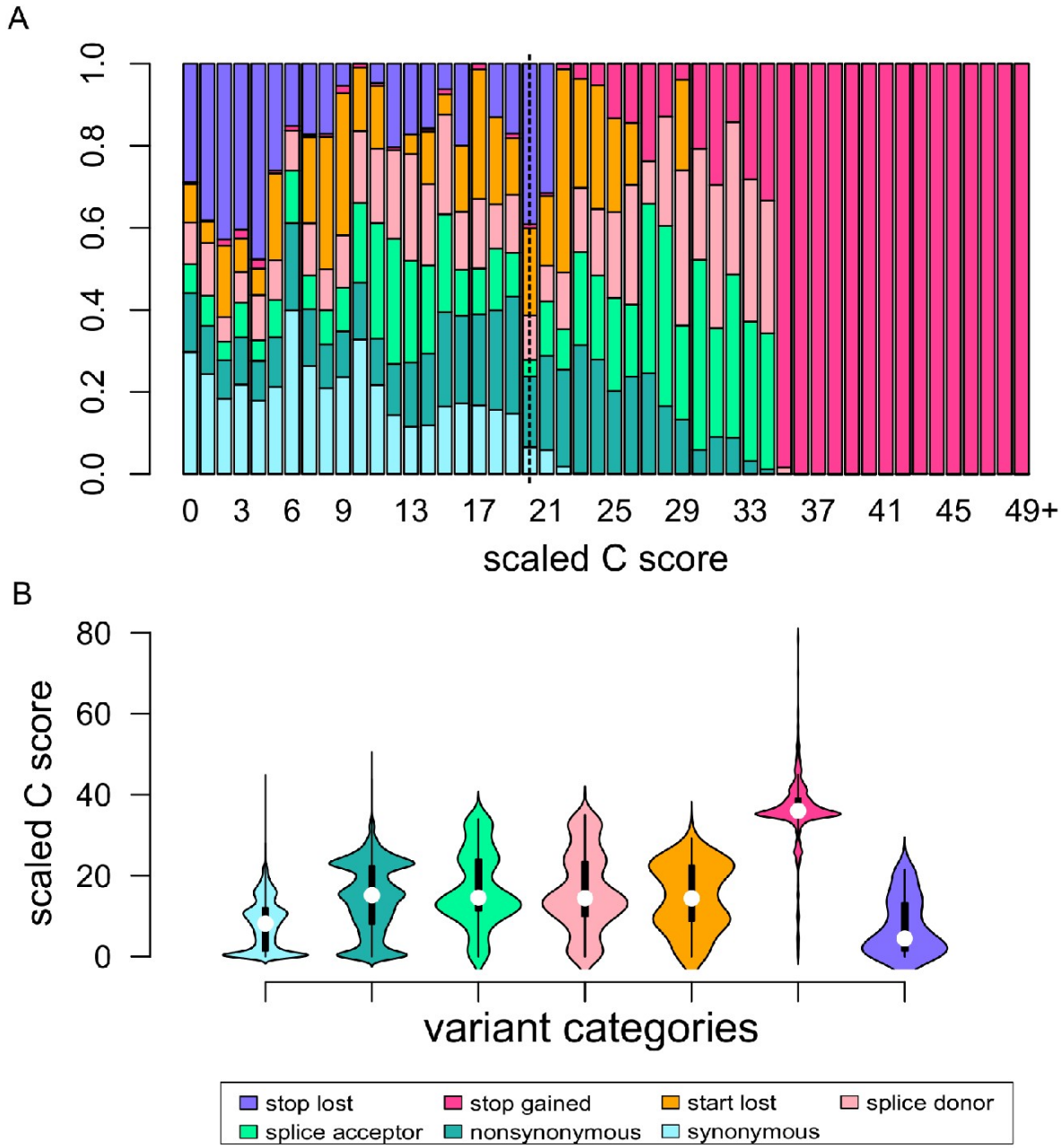


Figure S5.6: Relationship between scaled C scores and variant categories. (A) Proportions of each variant category per scaled C score, normalised by the total number of variants in each category. (B) Violin plots showing total distribution of scaled C scores for each variant category.

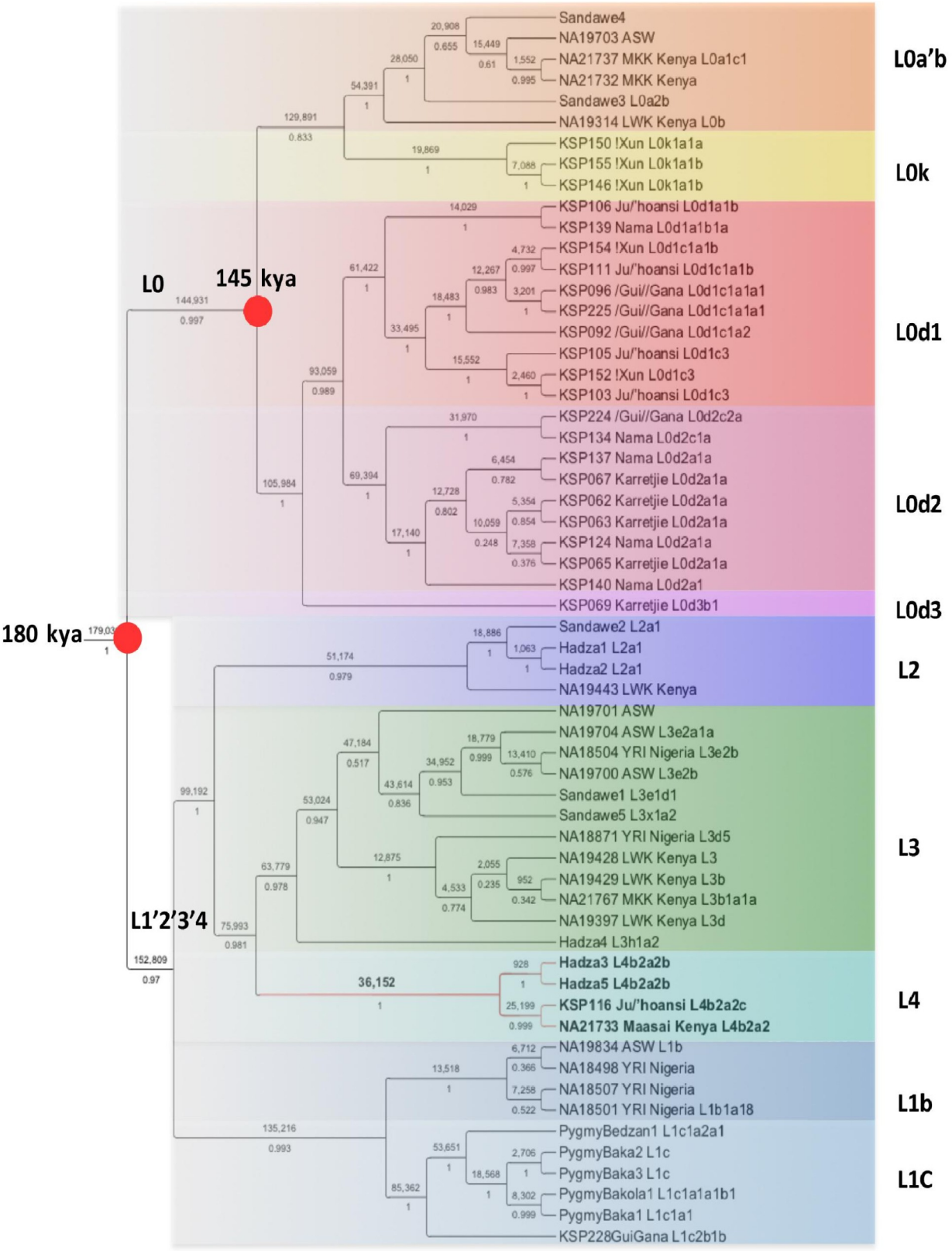


Figure S5.7: mtDNA tree, based on combined data including twelve populations of African origin. Clades that represent different sub-haplogroups are shaded with different colors. The numbers above the branches are node ages (median) and below are posteriors. Branch lengths are not scaled to time.

6. Population structure and admixture analysis

Several SNP genotype studies have focused on African populations (Patin et al. 2009; Henn et al. 2011; Pagani et al. 2012; Pickrell et al. 2012; Schlebusch et al. 2012; Triska et al. 2015; Busby et al. 2016; Hollfelder et al. 2017; Patin et al. 2017), demonstrating, for instance, that the Khoe-San were the most divergent human population. These studies also pinpointed the importance of migration, potentially driven by new ways of life, and its major impact on the distribution of current-day Africans. A prominent example is the 'Bantu-expansion', which led to people of west African descent spreading across most of sub-Saharan Africa (Schlebusch and Jakobsson 2018). The genetic variation represented on SNP genotype arrays, however, suffers from ascertainment bias; focusing only on known variants typically ascertained in a few (often) non-African-centered panels, which makes direct inferences problematic (Albrechtsen et al. 2010; Lachance and Tishkoff 2013). With the increasing availability of whole genome data from diverse sets of populations from Africa, we now have an opportunity to investigate African population structure in a more thorough and unbiased manner.

6.1. Principal Component Analysis

Principal Component Analysis (PCA) was performed on the “global dataset” (after exclusion of the two Hadza duplicates and of the two Khoe-San samples with lower quality) with EIGENSOFT (Patterson et al. 2006; Price et al. 2006) with the following parameters: no outlier removal, population size limit of 20.

The first two principal components (PC’s) of a globally representative dataset (27 million SNPs in 31 populations, Figure S6.1) explain 7.515% of the genetic variation in the dataset. PC1 is mainly determined by large dissimilarities between non-African and Khoe-San populations. PC2 is mainly determined by dissimilarities between Khoe-San and other African populations (the most extreme is the Mbuti sample). Together, PC1 and PC2 divide global genetic variation into three groups, non-African, Khoe-San and other Africans. Subsequent PC’s (PC 3, 4 and 6) summarize variation present in other African hunter-gatherer groups (eastern rain-forest hunter-gatherers, western rain-forest hunter-gatherers and Hadza) as well as variation within the Khoe-San group (northern vs southern vs central Khoe San, PC5 and PC7). Further structure and shared variation in Khoe-San groups are represented by PCs 9, 10 and 11. This PCA on a globally representative, full genomic dataset illustrates the extent of African diversity and is a true reflection of global genetic diversity, in contrast to SNP chip based PCA (for example (Schlebusch et al. 2012)), where non-African variation is amplified through the effects of ascertainment bias.

Initial SNP studies on the Khoe-San groups suggested various amounts of western African, European and eastern African admixture in the five Khoe-San groups included in this study (Schlebusch et al. 2012)). Indications of this admixture can also be seen on PC’s 1, 2 and 6 (Figure S6.1). It appears that !Xun and |Gui and ||Gana have comparatively more admixture from western African Bantu-speakers (PC2 and PC6). Nama have admixture from non-Africans (PC1) and some admixture from “other Africans” (PC2) which could be best interpreted as eastern African admixture (compare PC6 vs PC4). The Ju|'hoansi is the least admixed group while the Karretjie have two relatively un-admixed individuals and three individuals with more admixture (comparable to levels in !Xun, |Gui and ||Gana and Nama).

We also performed a PCA on the “extended global dataset” (preparation described in sub-section 3.9 - we excluded the same samples as in the first PCA). This dataset includes one ancient San sample (Ballito Bay A) and five additional modern Khoe-San samples (three Ju|'hoansi and two ǂKhomani) from the SGDP dataset. PC1 through PC6 are overall similar to the “global dataset” (Figure S6.2). From PC7 (not shown) axes are driven by dataset bias. The SGDP Ju|'hoansi samples cluster together with the KSP and HGDP Ju|'hoansi. The SGDP ǂKhomani samples cluster with the southern San (PC5). The ancient sample clusters within the Khoe-San diversity on PC1 and 2. On PC5 - which differentiates southern and northern San - it clusters at the extreme of the southern San (as observed in (Schlebusch et al. 2017)). We performed the same analysis but projected Ballito Bay

A on the modern-day diversity (Figure S6.3). The results are extremely similar. In this PCA, one of the two ‡Khomani samples is at the extreme of the northern-southern San PC (PC5).

6.2. Cluster analysis

ADMIXTURE (Alexander et al. 2009) analyses were performed in order to cluster individuals based on the genome sequence data and by extracting a set of SNPs. The global dataset was used for admixture analysis. We performed additional filtering with plink v.1.90b4.9: removal of SNPs with more than 0.1% missing data, removal of the two Hadza duplicates and the two KSP individuals with lower quality data, minor allele frequency threshold of 10%. SNPs in linkage disequilibrium were pruned based on pairwise genotypic variation using a window size of 50 SNPs shifting by 10 SNPs and a r^2 threshold of 0.7. The dataset decreased from 26,623,507 to 1,085,831 variants. Default settings and a random seed were used. Up to nine values of K (the number of clusters) were tested (K=2 to K=10). A total of 50 iterations of ADMIXTURE were run for each value of K. The iterations were analyzed and plotted with pong (Behr et al. 2016) with the greedy algorithm (Figure S6.4).

Admixture analysis recapitulated global population structure and also captured the northern vs. southern San split at K=9. Evidence of admixture into the Khoe-San, particularly into the Karretjie and the Nama, from an Eurasian/eastern African source is visible from K=2 onwards. Evidence of admixture from western African populations into the Khoe-San is also visible from K=3. The Hadza form their own cluster at K=5. The rain-forest hunter-gatherers form their own cluster at K=8.

We also performed an ADMIXTURE analysis on the “Extended global dataset” with similar filtering criteria: removal of SNPs with more than 0.1% missing data, removal of the two Hadza duplicates and the two KSP individuals with lower quality data, minor allele frequency threshold of 10%, LD pruning with a window size of 50 SNPs shifting by 10 SNPs and a r^2 threshold of 0.6. The dataset decreased to 1,161,558 SNPs. The clustering (Figure S6.5) is similar to that for the “global” dataset. The northern and southern Khoe-San split is captured at K=7 and the ancient sample has a majority from the southern Khoe-San component. The two SGDP ‡Khomani samples also have the majority of their ancestry in the southern San component while the three additional Ju|'hoansi individuals from the SGDP dataset belong to the northern San ancestry cluster.

6.3. Population trees with admixture edges

TreeMix analysis (Pickrell and Pritchard 2012) analysis was run on various combinations of populations allowing for a number of different admixture edges. The ancestry graph was rooted using all Khoe-San groups and standard errors were computed based on blocks of 2,000 SNPs. Figure S6.6 shows a tree with three admixture edges and the residual matrix of the fit, while Figure S6.7 shows trees with allowed admixture edges from one to ten.

With one allowed admixture edge, European and/or eastern African admixture (13.9%) in the Nama is inferred (Figure S6.7). This admixture might be explained partly as recent admixture with incoming European colonists and partly by interactions with a group that introduced herding practices in southern Africa (Pickrell et al. 2012; Breton et al. 2014; Macholdt et al. 2014) - see sub-section 6.7 for further discussion. The source population according to Treemix analysis is in-between eastern Africans and Europeans.

Two allowed edges indicates a strong admixture event between the eastern African Hadza group and the whole Khoe-San group (Figure S6.7). In the present TreeMix analysis the direction of admixture is indicated from Hadza into all Khoe-San groups, while previously (Pickrell et al. 2012), it was indicated in the opposite direction. The admixture fraction, however in the present study is large (49.1%), larger than found by (Pickrell et al. 2012) (23%). Since the admixture fraction inferred by our study is very close to 50%, the direction of admixture is uncertain and it is uncertain if such an event can even be considered as admixture (see sub-section 6.7 for further discussion). Possible other explanations of this high admixture component is that it is indicative of a combination of other

shared ancestries - when the Hadza is removed from the TreeMix analysis, a weaker eastern African admixture event into the base of the Khoe-San clade is inferred.

With three admixture edges, the !Xun and Ju|'hoansi have evidence of admixture (20.3%) with what appears to be a group of western African origin. The analysis suggest that the admixture came from a group in-between western rain-forest hunter-gatherer populations and western Africans (see sub-section 6.7 for further discussion on admixture in the !Xun).

At three allowed edges, the residual matrix indicates a good fit generally (Figure S6.6), with two possible constraints on the tree in that the San and Ju|'hoansi was not grouped on the tree and the Mbuti and Hadza seem to have some kind of association not represented by the tree.

With increasing number of edges, more possible admixture events between comparative groups become visible. The Nama - European/eastern African edge and the Hadza/Khoe-San edge remain constant throughout the analysis. Interestingly the !Xun (and the associated western rain-forest hunter-gatherer/ western African edge) change position with increasing number of admixture edges. The !Xun move to a more basal position in the Khoe-San tree (closer to rain-forest hunter-gatherer/ western African groups) and an admixture event from Ju|'hoansi into the !Xun is inferred instead.

6.4. D- and f_4-tests to investigate genetic connections to the Neandertal and Denisovan individuals

6.4.1. Testing admixture with Neandertal and Denisovan

We tested potential admixture with Neandertal and Denisovan by performing D-tests (Green et al. 2010) (Figures S6.8-13). We counted the number of sites in each of the following configurations: 011, 101, 011, where 011 mean that P1 has the ancestral variant, P2 has the derived and P3 has the derived, and calculated the following ratio: (number of sites that are 011 - number of sites that are 101)/(number of sites that are 011 + number of sites that are 101). As previously reported, we detect a strong Neandertal component in all non African individuals and a Denisovan component in the Papuan individual.

6.4.2. Testing differential introgression rates of Neandertals vs Denisovans

We tested the differential introgression rates of Neandertals versus Denisovans using the D-test with P1=Neandertal, P2=Denisovan and P3 a modern human individual (Figure S6.14). No differential affinity to the Neandertal vs Denisovan individuals was found across African individuals. All non African individuals except the Papuan show more introgressed Neandertal derived genetic material than Denisovan derived genetic material. A similar test using f_4=(Khoe-San - non-Khoe-San) (Denisovan - Neandertal) reveals the same pattern (Figure S6.15).

6.5. Setting Khoe-San as reference population (P3)

A negative D suggests more shared drift between P1 and P3 than between P2 and P3 and a positive D suggest more shared drift between P2 and P3 than between P1 and P3. P3 is in this set up a Khoe-San individual and thus P1 and P2 is always the closest phylogenetic relationship.

Figure S6.16 illustrates that while all Khoe-San individuals share more drift with non-Africans than with either rain-forest hunter-gatherers and (at least to some extent) western Africans, only the Nama and three of the five Karretjie individuals share more drift with non-Africans than with eastern Africans (top panel). The last panel is consistent with a more European than an Asian source, especially in Nama.

In Figures S6.17 and S6.18, P3 is also Khoe-San but here we more specifically investigate the relative affinity to the three eastern African populations Hadza, Sandawe and Maasai (MKK).

Figure S6.17 reveals that for all the Khoe-San populations, the Hadza, Sandawe and MKK share much more drift with the Khoe-San population than with rain-forest hunter-gatherer populations and also the western African populations but about the same amount as the Dinka (another eastern

African population). The tendency is the same for the Hadza, Sandawe and MKK individuals when using Karretjie, |Gui and ||Gana or !Xun individuals as the outgroup (P3) but not when using the Ju|'hoansi or Nama individuals as outgroup: there is a clear tendency for MKK individuals to share more drift with the Nama individuals than with either Sandawe or Hadza individuals and for the Hadza individual to share more drift with the Ju|'hoansi individuals than the MKK and Sandawe individuals do.

Contrasting to non-Africans instead of other Africans (Figure S6.18) and neglecting the Papuans, all the Nama and some of the Karretjie individuals share more drift with non-Africans than with eastern Africans. The other Karretjie individuals, the |Gui and ||Gana and the !Xun appear to share about the same drift with eastern Africans and non-Africans. The Ju|'hoansi on the other hand appear to share more drift with Hadza than with either Sandawe or Maasai.

The direction of the Hadza admixture is not clear: on the one hand, if the direction was from Hadza into Ju|'hoansi (not Ju|'hoansi into Hadza), the effect would only be visible in Ju|'hoansi -- consistent with Figure S6.17 -- on the other hand, this direction of admixture would suggest that all eastern African populations would share more drift with Ju|'hoansi that with non-Africans which is not apparent in Figure S6.18.

The affinity between Ju|'hoansi and Hadza was further investigated using f_4=(KSP1-KSP2)(Hadza-nonKSP) (Figure S6.19). A positive value of this statistic suggests affinity between either KSP1 and Hadza or KSP2 and X while a negative values suggests affinity between either KSP1 and X or KSP2 and Hadza. Assuming that Hadza only has affinity to Ju|'hoansi/San, these patterns reflect several events: 1) a relatively large component from a Bantu speaking source in !Xun and |Gui and || Gana, 2) a large European component in Nama, 3) a smaller, but still substantial, non-African component in Karretjie, 4) a smaller Bantu component in Mbuti compared to the other rain-forest hunter-gatherer populations.

6.6. Admixture dating

We used ADMIXTOOLS (Patterson et al. 2012) to estimate the linkage disequilibrium (LD) decay due to admixture and thereby infer dates of admixture (ROLLOFF analysis). Default parameters were used. Various ROLLOFF decay curves were estimated using different populations as the two parental reference populations (Table S6.1). The standard error was estimated with a jackknife procedure implemented in the ROLLOFF package. Results are summarized in Table S6.1 and LD decay curves are shown in Figures S6.20-23. Results are discussed in sub-section 6.7.

6.7 General discussion on Khoe-San population structure and admixture

Global population structure analysis clearly illustrates the deep divergence and high diversities of the Khoe-San populations (PCA: Figures S6.1-3, ADMIXTURE: Figures S6.4-5; Treemix: Figures S6.6-7). The first two PC's of a globally representative dataset (27 million SNPs in 31 populations) explain 7.515% of the genetic variation in the dataset and divide global genetic variation into three groups, non-African, Khoe-San and other Africans. Subsequent PC's summarize variation present in other African hunter-gatherer groups (eastern rain-forest hunter-gatherers, western rain-forest hunter-gatherers and Hadza) as well as variation within the Khoe-San group (northern vs southern vs central Khoe San).

Initial SNP studies on the Khoe-San groups suggested various amounts of western African, European and eastern African admixture in the five Khoe-San groups included in this study (Schlebusch et al. 2012). Suggestions of this admixture can also be seen on PC's 1, 2 and 6 (Figure S6.1). We further investigated these possible admixture events by doing D-tests and *f*-statistics (Figures S6.8-13 and S6.16-19), looking at pairwise shared private alleles (Main Figure 2), and performing Treemix analysis (Figures S6.6-7); we thereafter dated (Patterson et al. 2012) the admixture events (Table S6.1, Figures S6.20-23).

We found clear evidence of recent European and/or eastern-African admixture (13.9%) in the Nama (Figures 2, S6.6-7, S6.8-13, S6.16-19). This might be explained partly as recent admixture with incoming European colonists. However the admixture event is dated to well before the arrival of colonists (Table S6.1, Figures S6.20) and coincides with the first evidence of herding practices in southern Africa, possibly introduced by an external group (Breton et al. 2014; Macholdt et al. 2014; Ranciaro et al. 2014). The source population is difficult to determine, both eastern African and European sources are likely (Figures 2, S6.6-13, S6.16-S6.20: CEU_Juhoansi_Nama + MKK_Juhoansi_Nama) as was discussed in previous studies. The eastern African and European components are difficult to distinguish and it is possible that we do not have a good representative parental population in the dataset.

We furthermore found strong evidence of an admixture event between the eastern African Hadza group and the whole Khoe-San group (Figures S6.6-7). This evidence is the clearest in the Ju|'hoansi (D-tests - Figures S6.15-18) and was previously detected in SNP studies (Pickrell et al. 2012; Schlebusch et al. 2012). mtDNA and Y-chromosome evidence (see sub-section 5.10.3) support the admixture event. In the present Treemix analysis the direction of admixture is indicated from Hadza into all Khoe-San groups, while previously (Pickrell et al. 2012), it was indicated in the opposite direction. The admixture fraction, however in the present study is huge (49.1%), larger than found by (Pickrell et al. 2012) (23%). Since the admixture fraction inferred by our study is very close to 50%, the direction of admixture is uncertain and it is uncertain if such an event can even be considered as admixture. When looking at admixture linkage disequilibrium decay (Table S6.1, Figure S6.20) of various San groups as one source and various eastern and western Africans as the other source, no evidence is found of admixture LD decay and an exponential fit was not significant. Thus a recent admixture event where Hadza is the recipient population is not likely. We also looked at the admixture LD where Khoe-San populations are the recipient population of Hadza admixture (Figure S6.23), while significant exponential fits were obtained here, dates correlate with eastern and western African admixture into the various Khoe-San groups and it is likely that these signals are indistinguishable (Table S6.1). A relatively recent admixture event between the Hadza and Khoe-San groups therefore seems unlikely. Possible other explanations of this high admixture component is that it is indicative of a combination of other shared ancestries - when the Hadza is removed from the Treemix analysis, a weaker eastern African admixture event into the base of the Khoe-San clade is inferred. This general eastern African component in all Khoe-San is also visible with negative D-tests in all Khoe-San (Figure S6.17-18). It could be that the observed admixture signal from the Hadza into the Khoe-San is a combination of eastern African admixture into all Khoe-San groups (linked with the herding introduction discussed above) combined with a deeper shared affinity of Hadza and Khoe-San populations due to isolation by distance (historically Hadza ancestors would have been the most proximate population to Khoe-San groups geographically, of the groups included in this study) or a much earlier event. It could be that Ju|'hoansi was not affected by the more recent event (herding) as much as the other groups, and therefore the Hadza admixture is more visible in the Ju|'hoansi compared to other Khoe-San groups.

The !Xun and |Gui and ||Gana have evidence of admixture (20.3%) with what appears to be a group with western African origin (Figure S6.1 and S6.4). From the Treemix analysis it appears that the admixture came from a group in-between rain-forest hunter-gatherer populations and western Africans. The possibility exists that the correct parental population is not present in the dataset (or that the group has no descendants alive today). Of the five Khoe-San groups studied here, the !Xun appears to share the most private alleles (Main Figure 2) with all external African groups tested (Mbuti, eastern and western Africans). Among the Khoe-San groups, they also had the highest heterozygosity and allelic diversity (Table S5.1, Figures S5.1 and S5.4). It seems that the !Xun received the highest external gene flow from other Africans compared to other Khoe-San groups. The largest proportion comes from Mbuti followed by western Africans and eastern Africans (Main Figure 2). This order seems to be followed by other Khoe-San groups, however, the Nama share more private alleles with eastern Africans than with western Africans and the Ju|'hoansi have very even amounts of eastern and western African sharing.

Looking at private alleles with Khoe-San groups from the comparative group perspective (Figure S5.5), the Nama share comparatively more private alleles with non-Africans; !Xun and |Gui and ||Gana more with western Africans; !Xun, |Gui and ||Gana and Nama with eastern Africans; and !Xun and Ju|'hoansi with Mbuti. Since the !Xun and Ju|'hoansi is geographically the most northern groups, their increased private allele sharing with rain-forest hunter-gatherers is not unexpected. The central African Mbuti rain-forest hunter-gatherer group seems to be a focal group, with high amounts of private allele sharing with most other African groups. The geographic area north of the Khoe-San groups and south of the rain-forest hunter-gatherer and Hadza groups, is today occupied by Bantu speakers that arrived relatively recently in the area (~5 kya, (Li et al. 2014)), and possibly replaced groups that perhaps were genetically intermediate to the Khoe-San and central and eastern African hunter-gatherers. It is possible that the northern Khoe-San groups such as the !Xun had more gene flow with these extinct groups, and that we are observing signals of this gene flow in our analysis (Figures S6.6-7, S6.5) albeit with sub-optimal comparative groups. Also, it is likely that Bantu speakers absorbed genetic components of these groups and subsequent admixture with the !Xun and |Gui and ||Gana introduced more of these components. From admixture LD decay, it seems that the admixture dates into the !Xun, |Gui and ||Gana and Karretjie People correlates with the arrival of Bantu-speakers in their respective areas. The Ju|'hoansi and Nama individuals included in our study seem to have had less Bantu-speaker admixture (Figures 2 and S6.8-13, S6.16-18).

Population structure and admixture Tables and Figures

Table S6.1: Admixture dates according to LD decay patterns.

Source1	Source2	Admixed	Mean_Generations	Residual-SE	SigCode[1]
Admixture of outside groups (CEU YRI MKK) into the various Khoe-San groups					
CEU	Ju|'hoansi	Karretjie	6.022357	0.003803	333
YRI	Ju|'hoansi	Karretjie	6.512567	0.004062	333
MKK	Ju|'hoansi	Karretjie	5.6836224	0.003688	333
CEU	Ju|'hoansi	Nama	50.757042	0.004423	333
YRI	Ju|'hoansi	Nama	52.842357	0.004012	333
MKK	Ju|'hoansi	Nama	55.029601	0.003873	333
CEU	Ju|'hoansi	|Gui||GhanaKgal	20.968759	0.00468	333
YRI	Ju|'hoansi	|Gui||GhanaKgal	16.272118	0.005145	333
MKK	Ju|'hoansi	|Gui||GhanaKgal	17.272757	0.004602	333
CEU	Ju|'hoansi	!Xun	31.150206	0.004683	133
YRI	Ju|'hoansi	!Xun	32.853272	0.00433	333
MKK	Ju|'hoansi	!Xun	37.110795	0.00393	333
Central San as admixed group between northern and southern San					
Karretjie	Ju|'hoansi	|Gui||GhanaKgal	11.103467	0.004274	333
Hadza admixture into San or San admixture into Hadza					
Admixture into Hadza with Ju|'hoansi/Karretjie as one source and western/eastern Africans as other source					
MKK	Ju|'hoansi	Hadza	160.35295	0.003445	101
MKK	Karretjie	Hadza	25.41494	0.003684	031
Sandawe	Ju|'hoansi	Hadza	123.3109	0.004316	000
Sandawe	Karretjie	Hadza	185.5005	0.004353	000
YRI	Ju|'hoansi	Hadza	226.38432	0.00448	300
YRI	Karretjie	Hadza	453.6644	0.004367	200
Hadza and Ju|'hoansi as sources and admixture into various other San groups					
Hadza	Ju|'hoansi	Karretjie	5.8604554	0.003993	333
Hadza	Ju|'hoansi	Nama	53.270741	0.003962	333

Hadza	Ju|'hoansi	|Gui||GhanaKgal	18.209407	0.004224	333
Hadza	Ju|'hoansi	!Xun	37.166219	0.003982	333
Hadza and Karretjie as sources and admixture into various other San groups					
Hadza	Karretjie	Nama	56.132616	0.003824	333
Hadza	Karretjie	|Gui||GhanaKgal	16.260836	0.00404	333
Hadza	Karretjie	Ju|'hoansi	103.66101	0.003516	033
Hadza	Karretjie	!Xun	42.685371	0.003895	333

1 SigCode - Significance code - the level of significance of the three terms “A”, “C” and “m” of the exponential fit given by the formula “wcorr ~ (C + A * exp(-m * dist/100))”. Where “3” is pval <= 0.001; “2” is pval <= 0.01; “1” is pval <= 0.05 and “0” is pval > 0.05. Thus “333” would mean that all three terms in the exponential fit formula had significant p-values of <0.001 while with “000” none of the terms were significant.

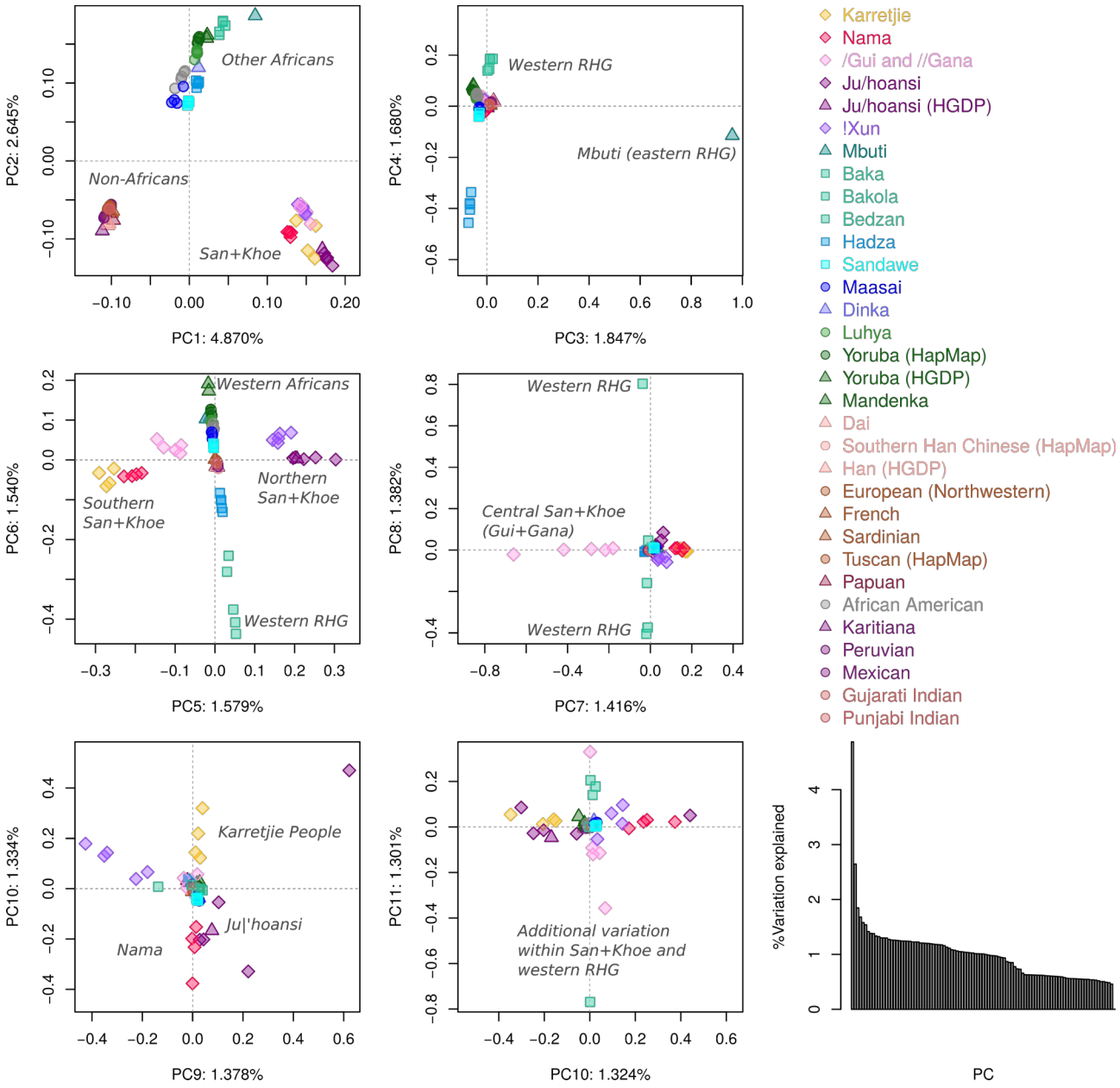


Figure S6.1: Principal component analysis of Global dataset. RHG: rain-forest hunter-gatherers.

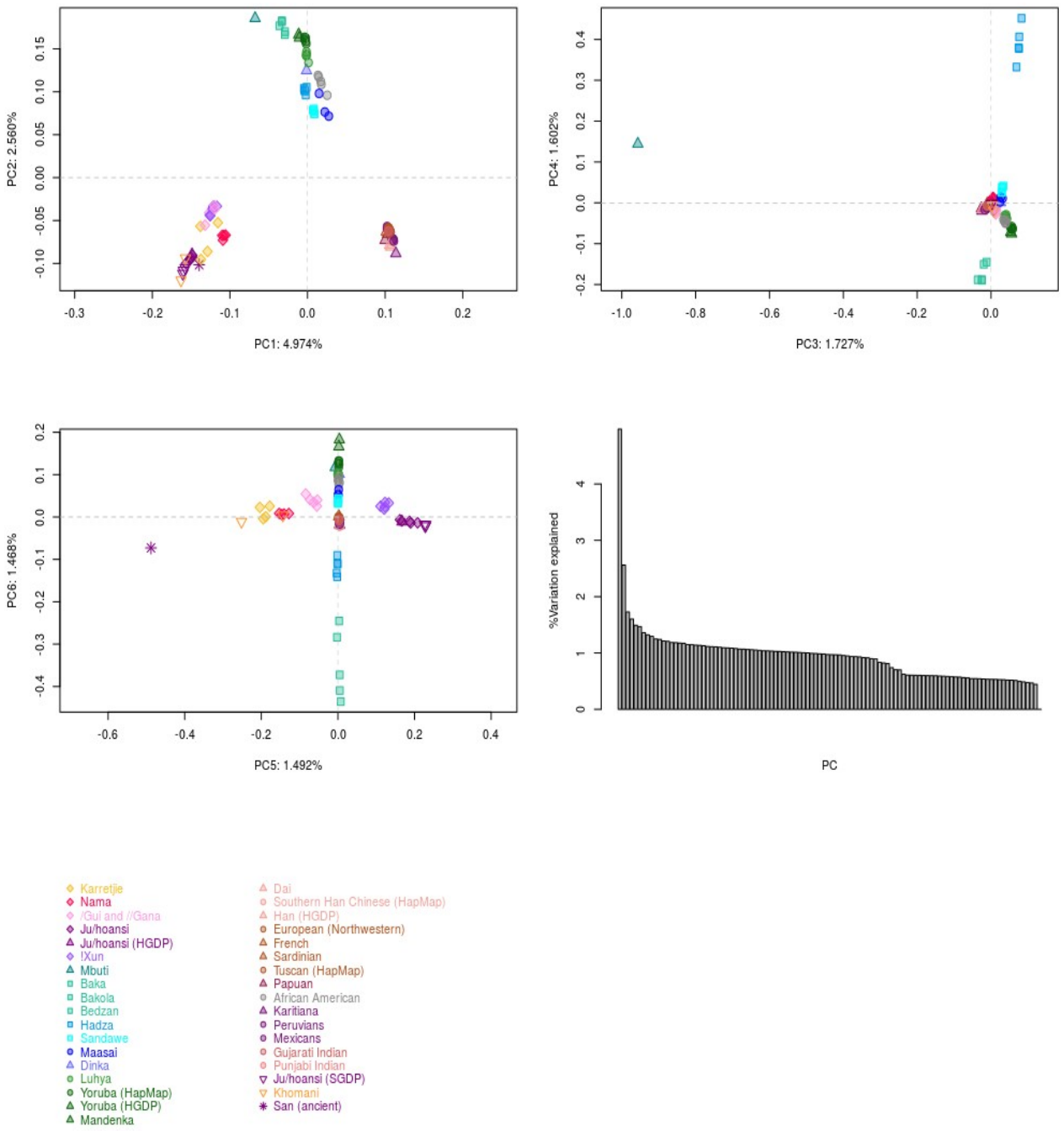


Figure S6.2: Principal component analysis of Extended Global dataset

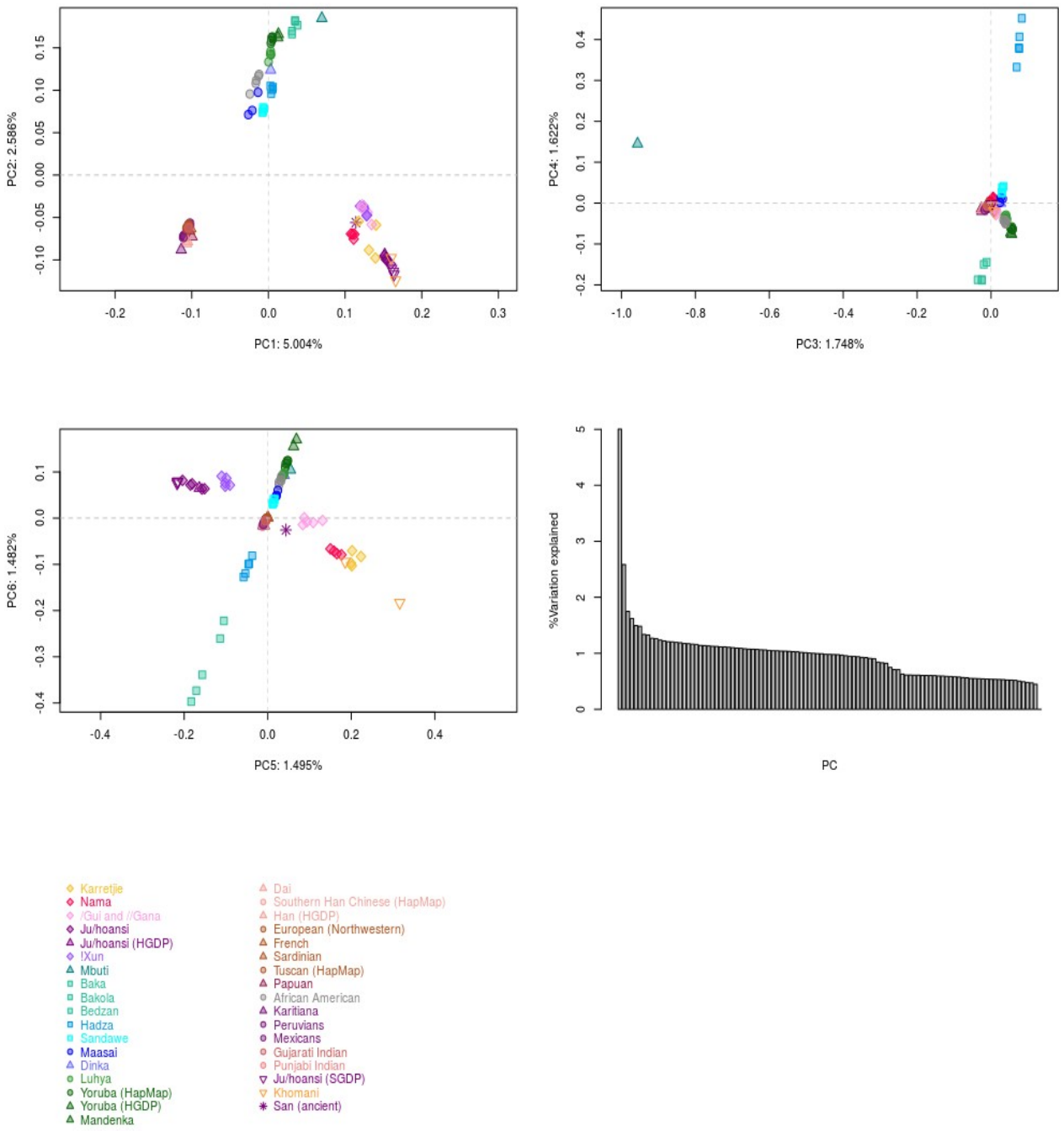


Figure S6.3: Principal component analysis of Extended Global dataset. The ancient San sample is projected.

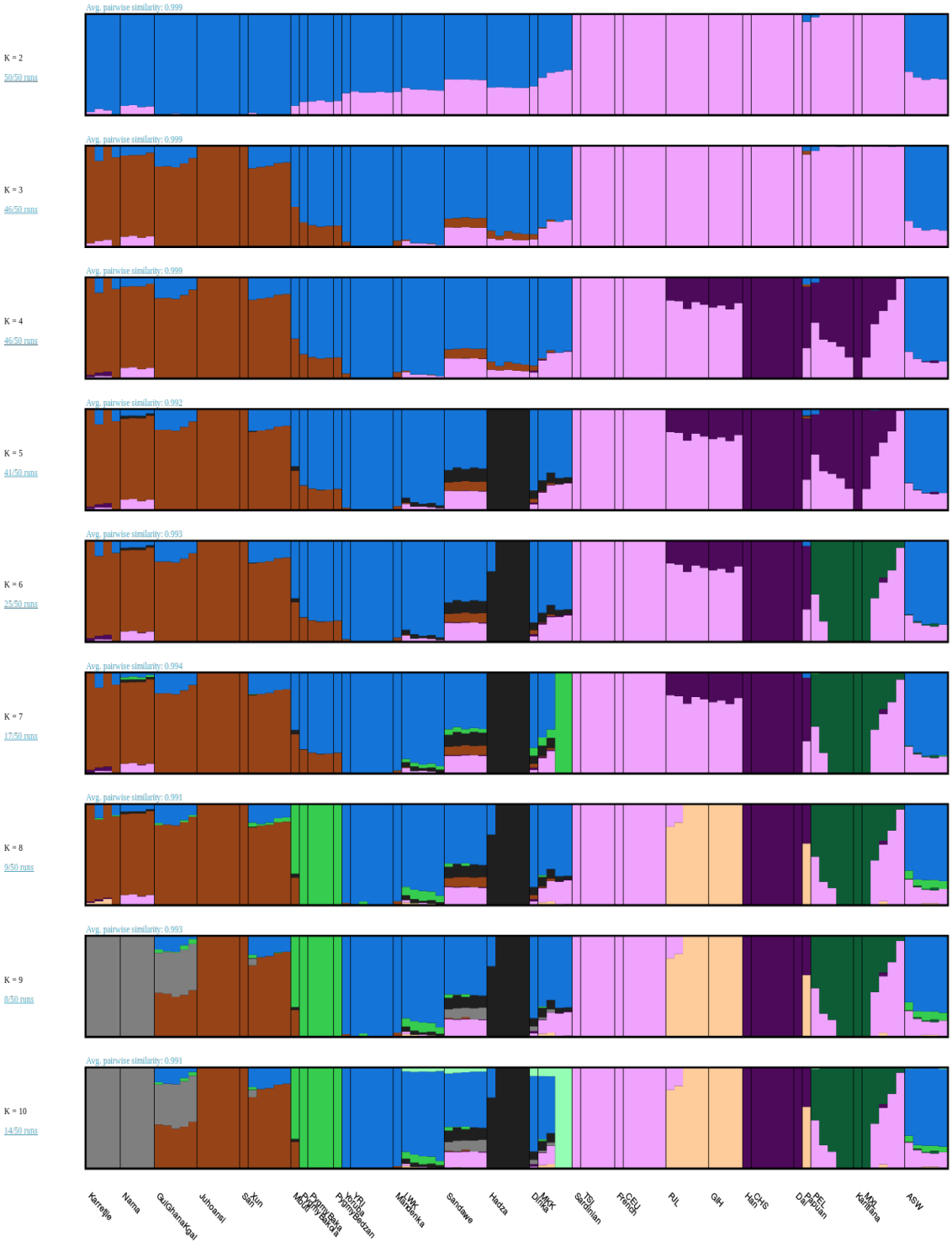


Figure S6.4: Admixture analysis of Global dataset. K=2 to K=10, 50 repeats.

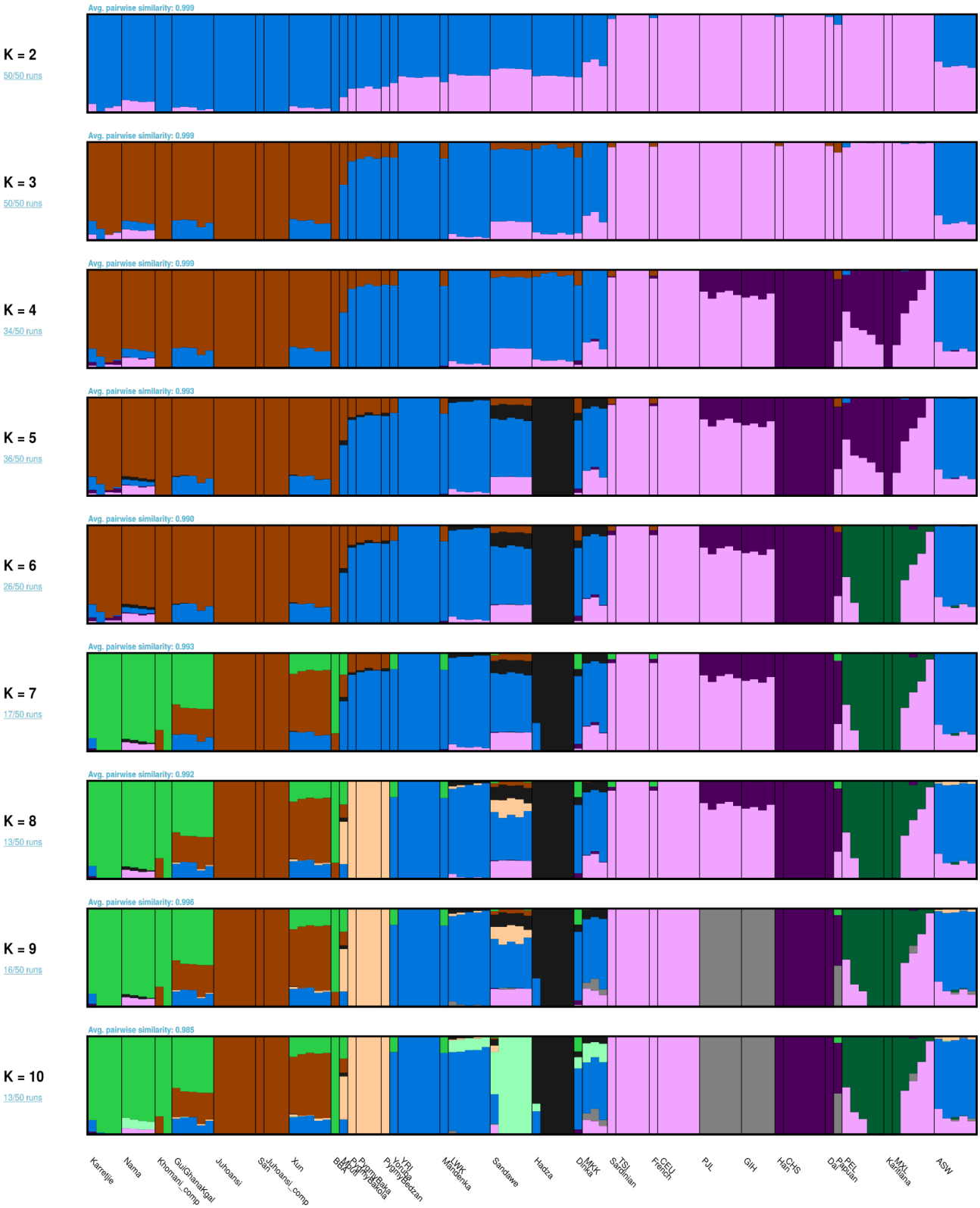


Figure S6.5: Admixture analysis of Extended Global dataset. K=2 to K=10, 50 repeats.

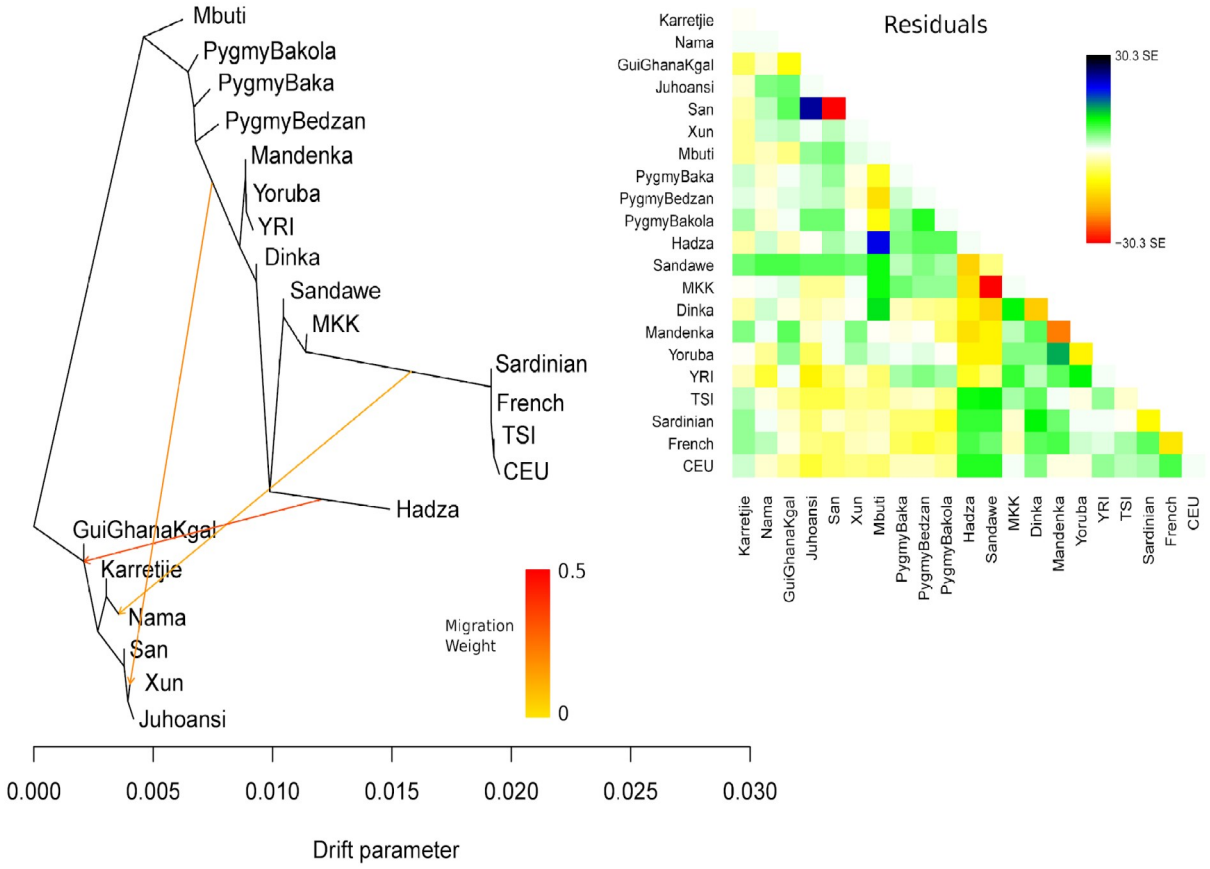


Figure S6.6: Treemix analysis of Global dataset with three admixture edges. (A) tree, (B) residual matrix of the fit.

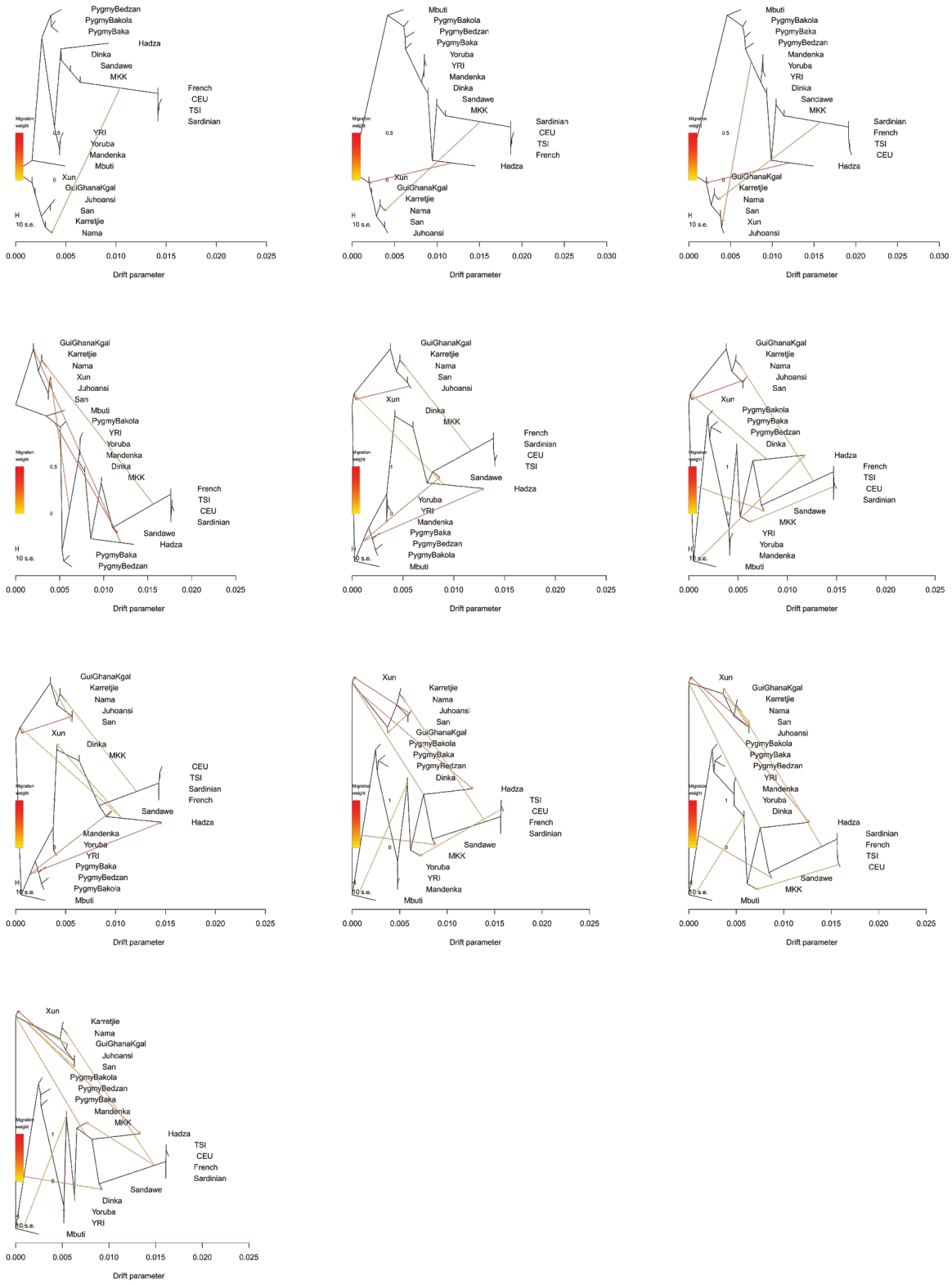


Figure S6.7: Treemix analysis of Global dataset, showing one to ten admixture edges.

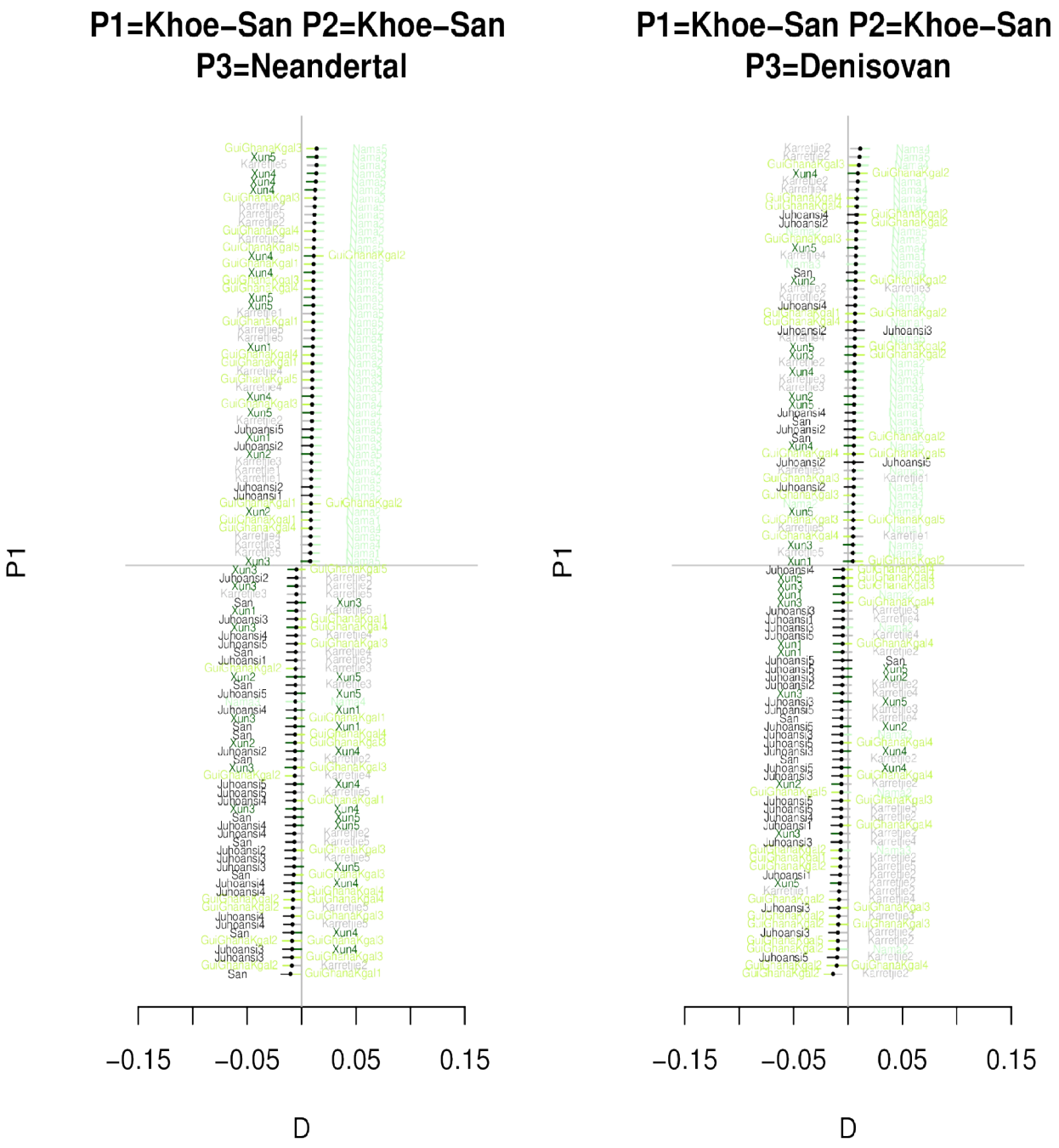


Figure S6.8: D-tests Global dataset with archaic humans (Neandertal and Denisovans) and Khoe-San. P1 and P2 are Khoe-San individuals and P3 is Neandertal (left graph) or Denisovan (right graph). Different colors correspond to the phylogenetic position of the population the individual was sampled from. The bars show ± 2 standard deviations based on a weighted block jackknife approach with 5 Mb windows. Only comparisons with the 50 smallest and the 50 largest values are shown.

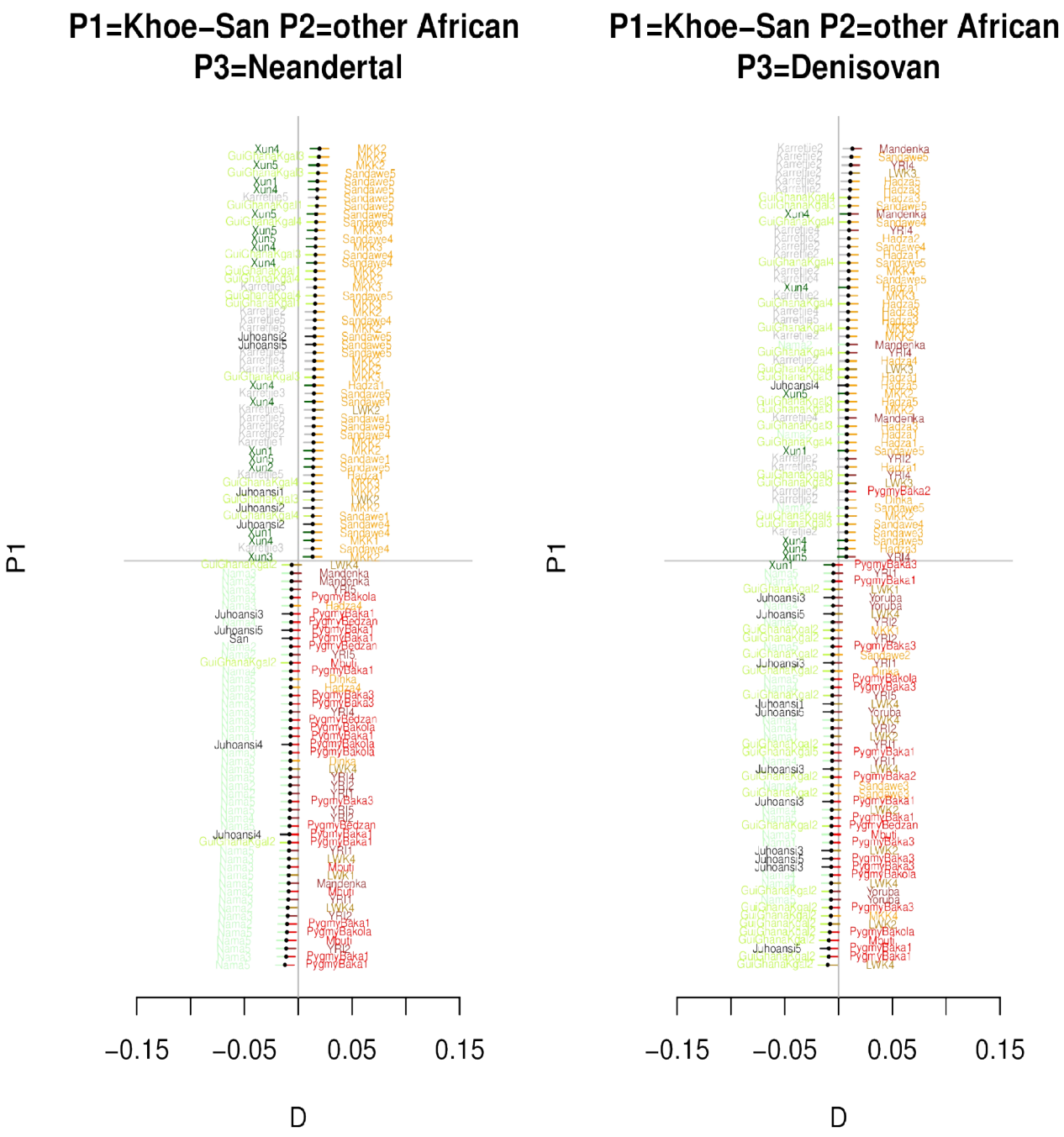


Figure S6.9: D-tests Global dataset with archaic humans (Neandertal and Denisovans), Khoe-San and Other Africans. P1 are Khoe-San individuals, P2 are African but not Khoe-San individuals and P3 is Neandertal (left graph) or Denisovan (right graph). Different colors correspond to the phylogenetic position of the population the individual was sampled from. The bars show ± 2 standard deviations based on a weighted block jackknife approach with 5 Mb windows. Only comparisons with the 50 smallest and the 50 largest values are shown.

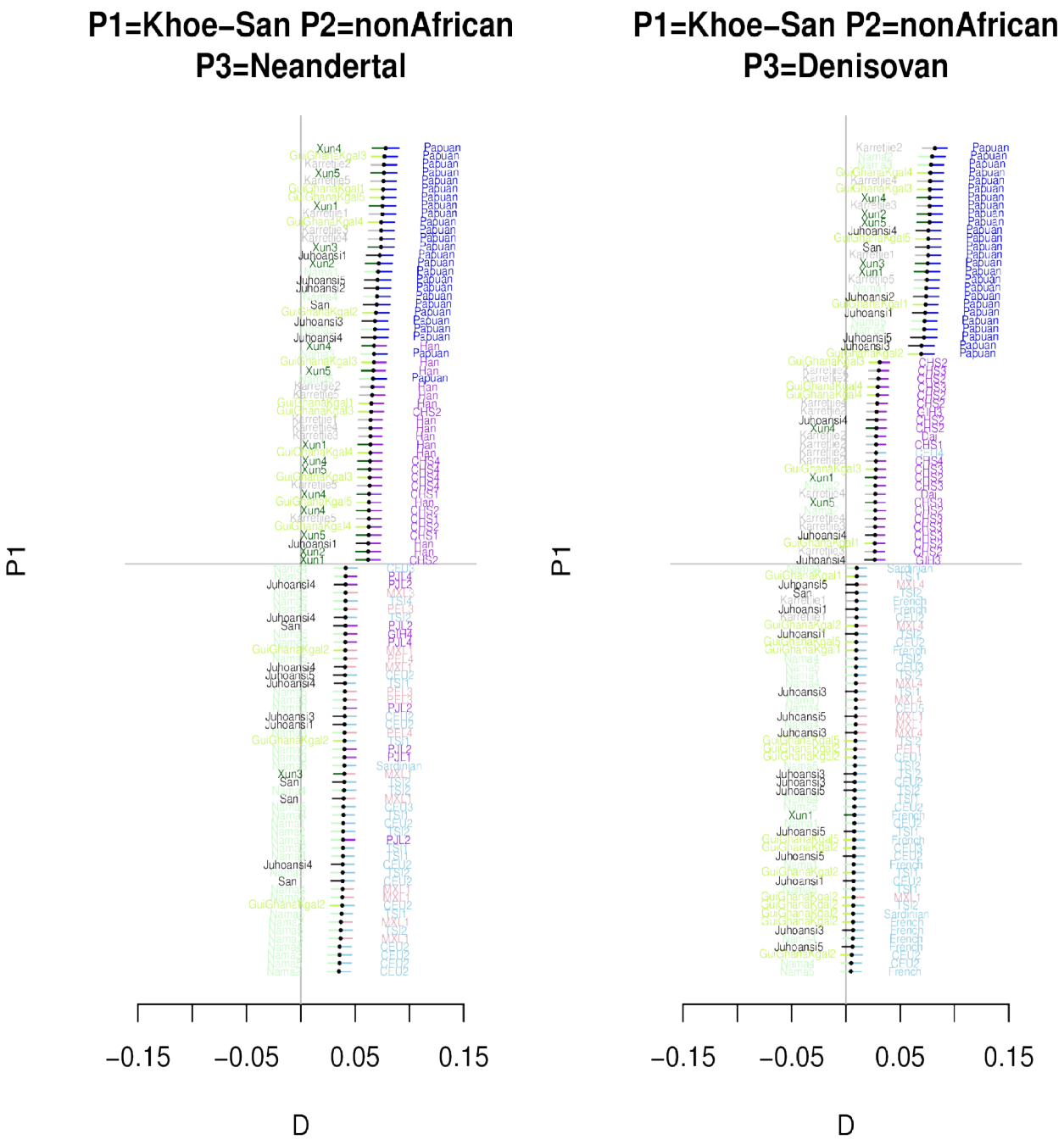


Figure S6.10: D-tests Global dataset with archaic humans (Neandertal and Denisovans), Khoe-San and non-Africans. P1 are Khoe-San individuals, P2 non-African individuals and P3 is Neandertal (left graph) or Denisovan (right graph). Different colors correspond to the phylogenetic position of the population the individual was sampled from. The bars show ± 2 standard deviations based on a weighted block jackknife approach with 5 Mb windows. Only comparisons with the 50 smallest and the 50 largest values are shown.

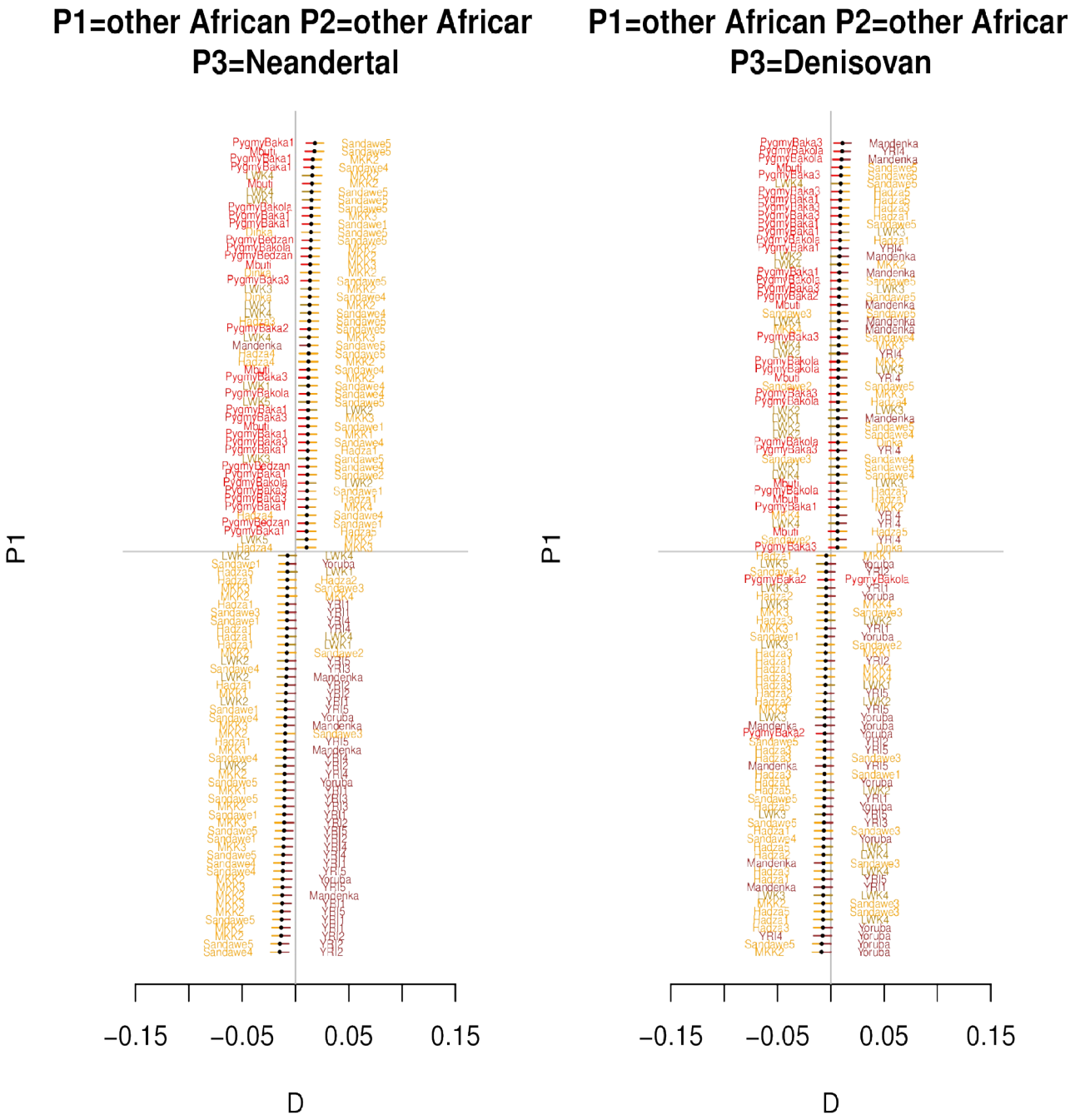


Figure S6.11: D-tests Global dataset with archaic humans (Neandertal and Denisovans) and other Africans. P1 and P2 are African but not Khoe-San individuals and P3 is Neandertal (left graph) or Denisovan (right graph). Different colors correspond to the phylogenetic position of the population the individual was sampled from. The bars show ± 2 standard deviations based on a weighted block jackknife approach with 5 Mb windows. Only comparisons with the 50 smallest and the 50 largest values are shown.

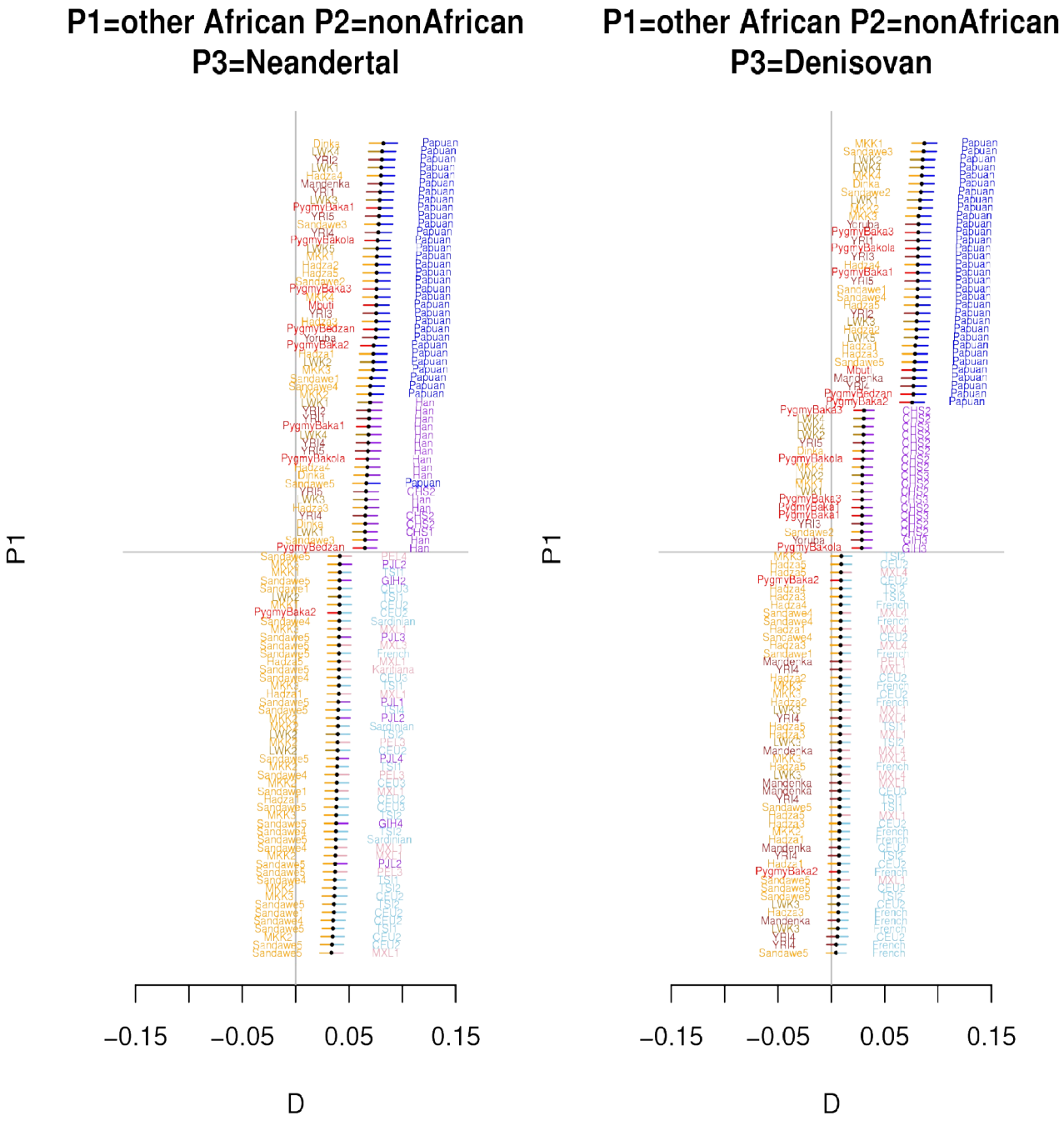


Figure S6.12: D-tests Global dataset with archaic humans (Neandertal and Denisovans), Other Africans and Non-Africans. P1 are African but not Khoe-San individuals, P2 are non-African individuals and P3 is Neandertal (left graph) or Denisovan (right graph). Different colors correspond to the phylogenetic position of the population the individual was sampled from. The bars show ± 2 standard deviations based on a weighted block jackknife approach with 5 Mb windows. Only comparisons with the 50 smallest and the 50 largest values are shown.

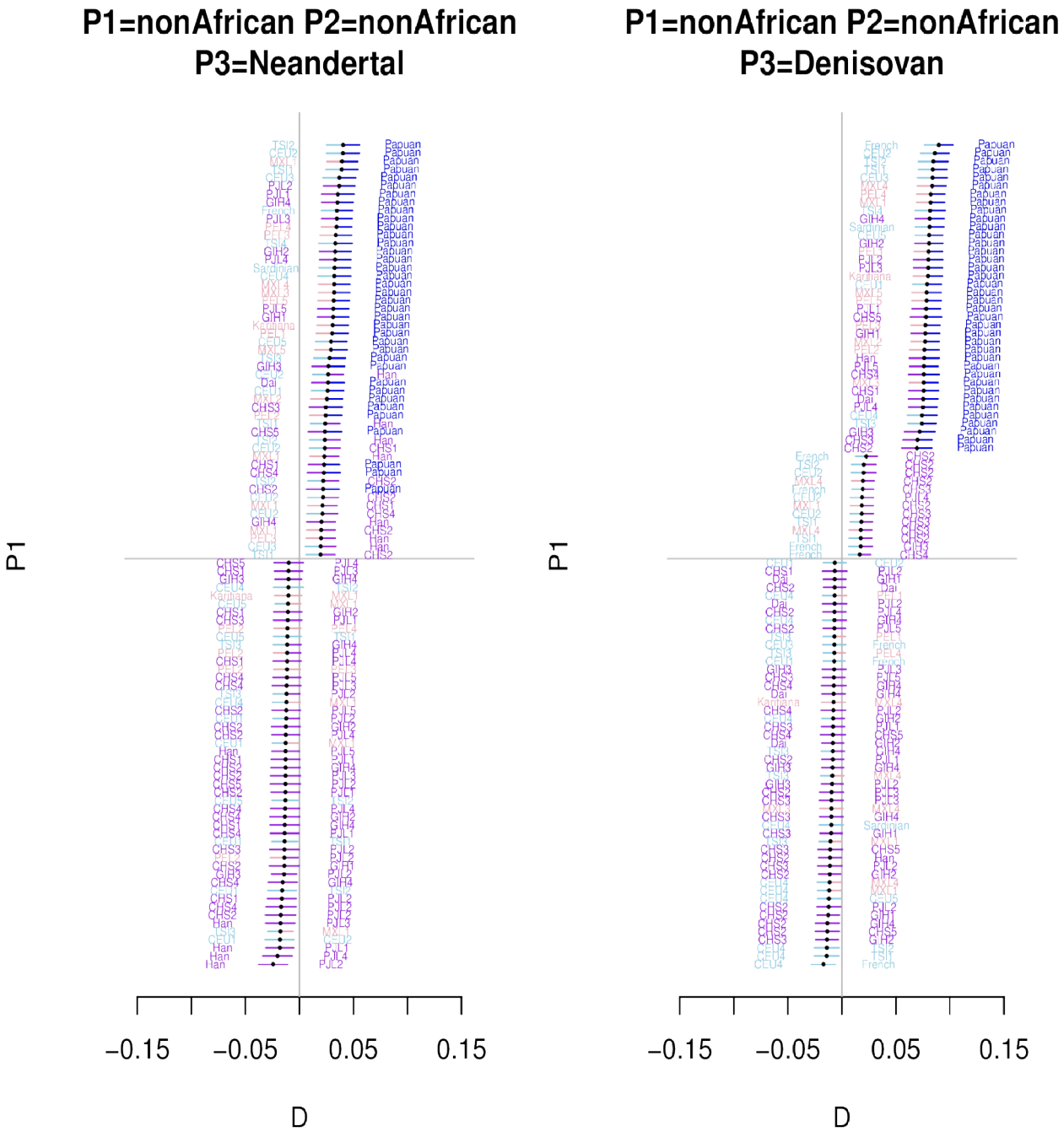


Figure S6.13: D-tests Global dataset with archaic humans (Neandertal and Denisovans) and non-Africans. P1 and P2 are non-African individuals and P3 is Neandertal (left graph) or Denisovan (right graph). Different colors correspond to the phylogenetic position of the population the individual was sampled from. The bars show ± 2 standard deviations based on a weighted block jackknife approach with 5 Mb windows. Only comparisons with the 50 smallest and the 50 largest values are shown.

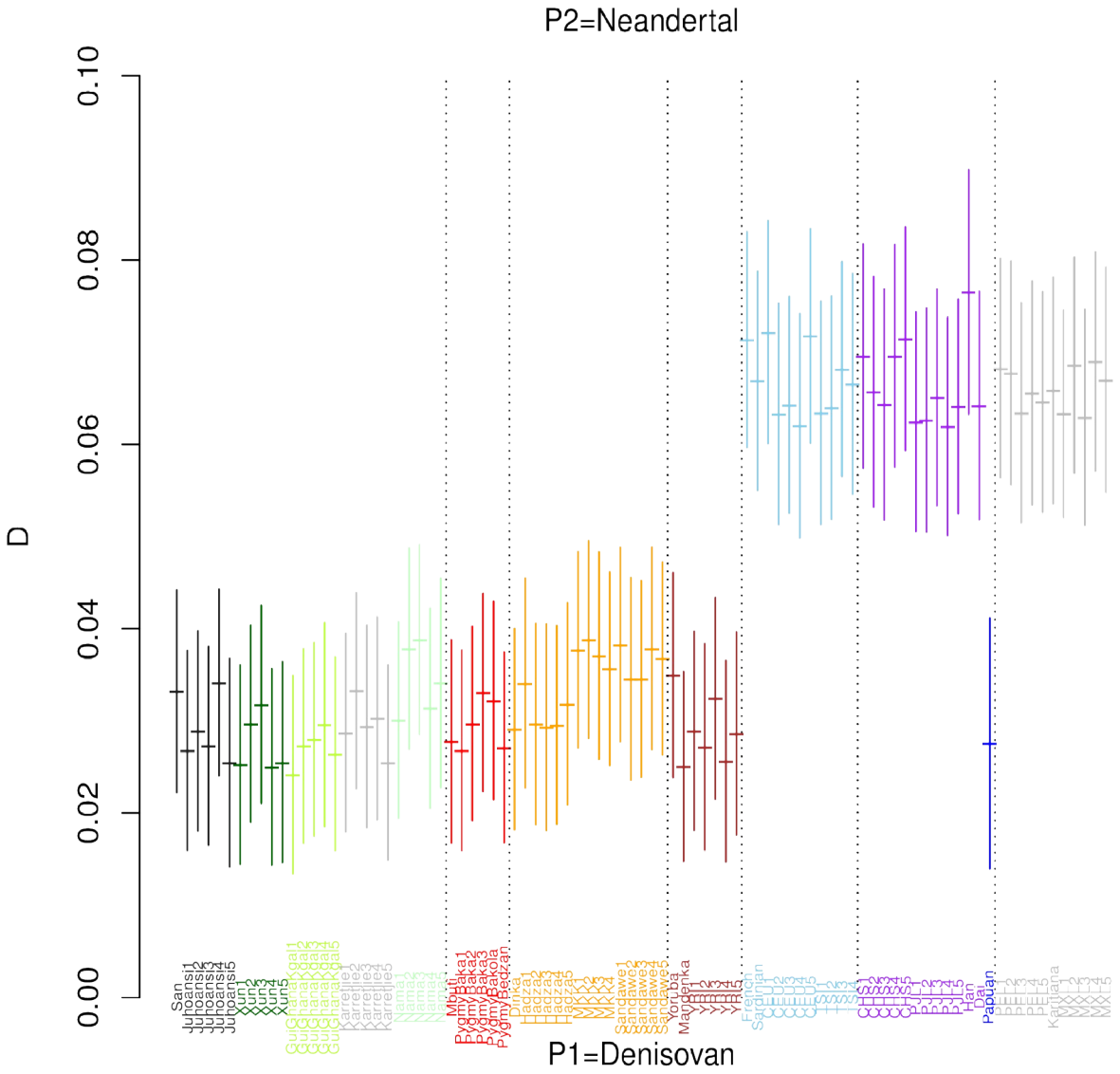


Figure S6.14: D tests - contrasting Neandertal and Denisovan. A positive D suggests more shared drift between P3 and Neandertal than between P3 and Denisovan. All investigated modern individuals share more drift with Neandertal than with Denisovan. There is remarkably little variation across African populations suggesting that any archaic admixture in African populations is either older than the Khoe-San branch split or involving an archaic human population equally related to the Denisovans and Neandertals.

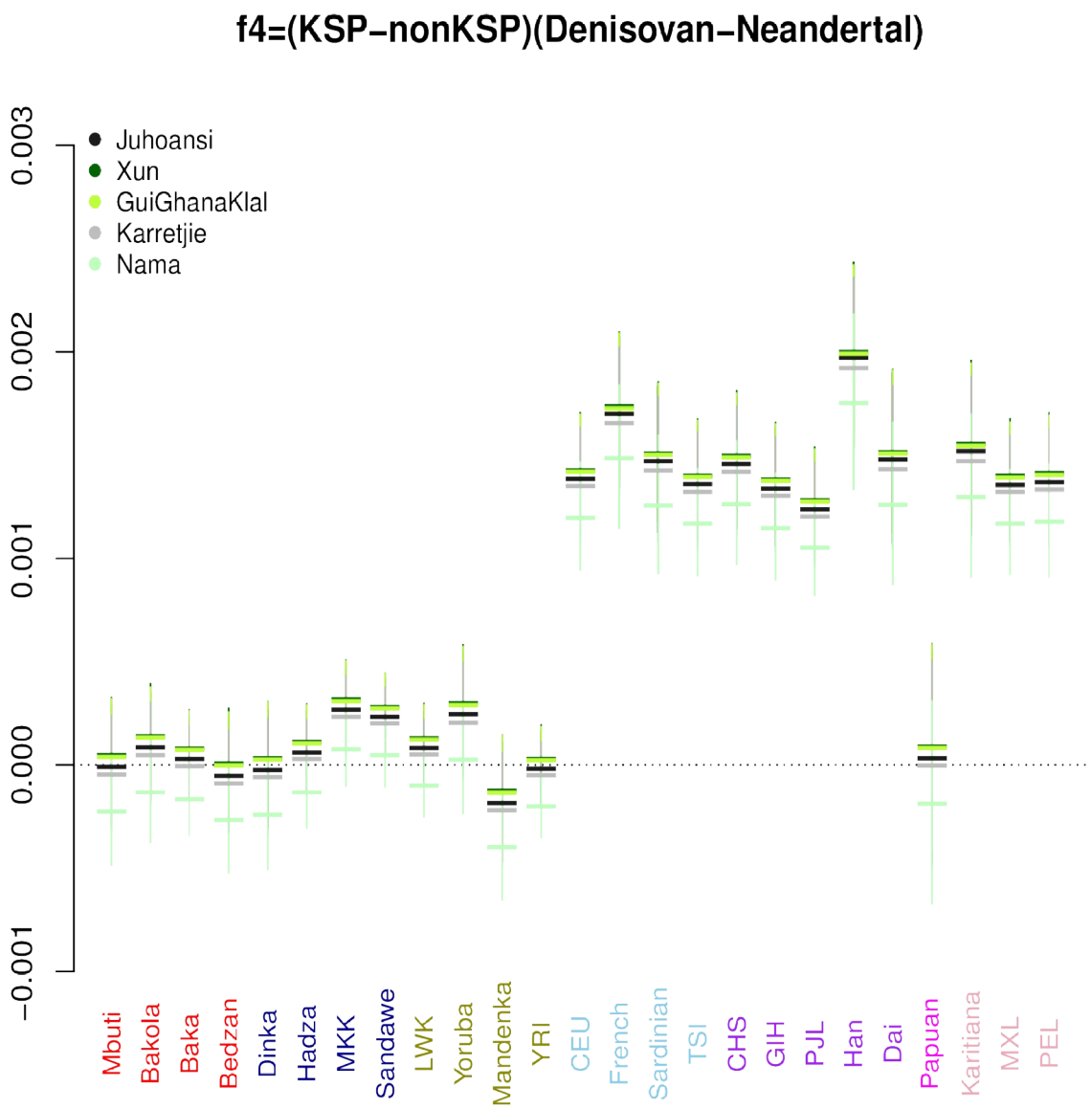


Figure S6.15: f_4 relative rate of Neandertal vs Denisovan admixture. A clear signals of more admixture with Neandertals than Denisovans in all the non-African samples except the Papuan individual. No clear such signal in the African samples.

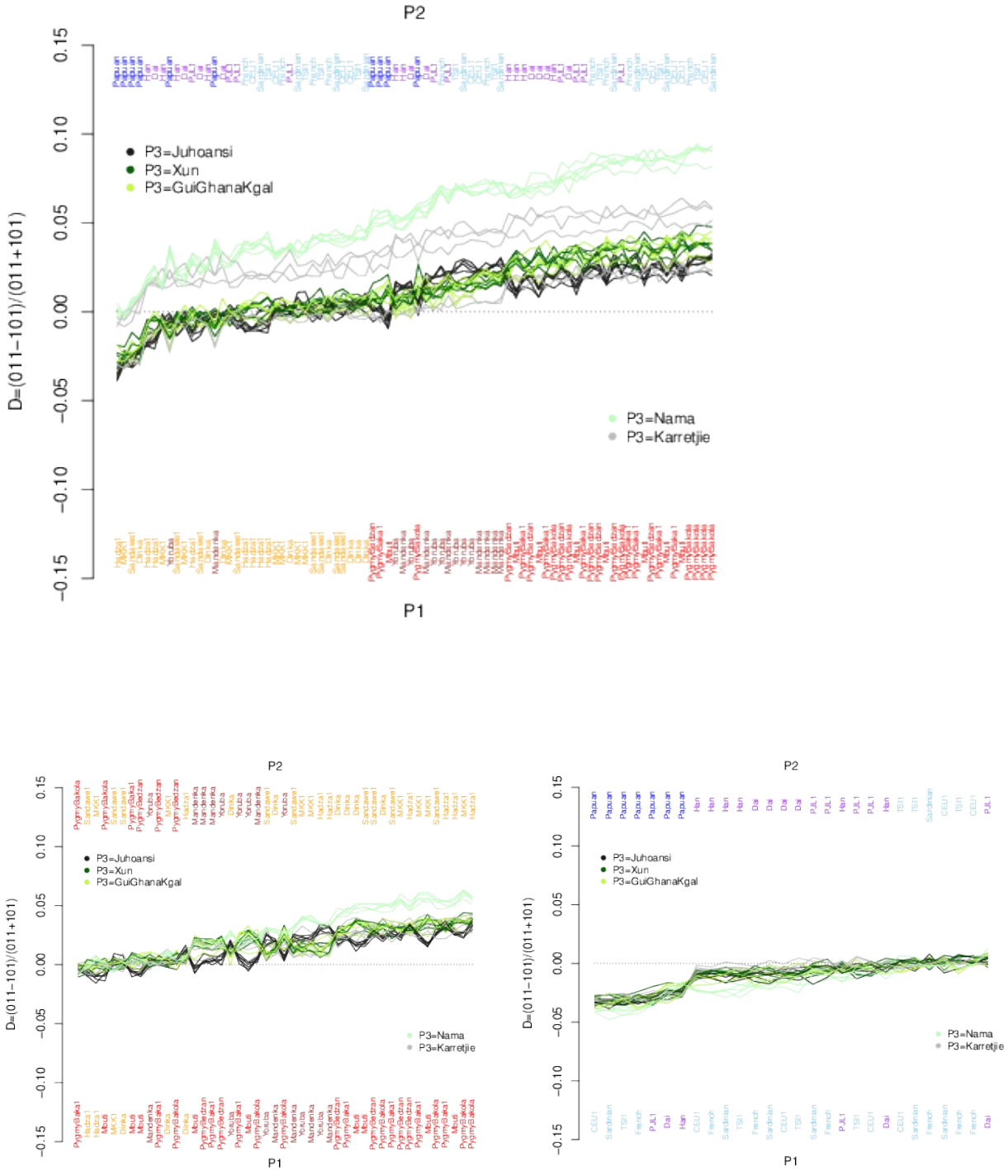


Figure S6.16: D-tests Global dataset. The top figure contrasts Africans and non-Africans. The bottom figure left contrasts African individuals. The bottom figure right contrasts non-Africans. Standard deviations not shown.

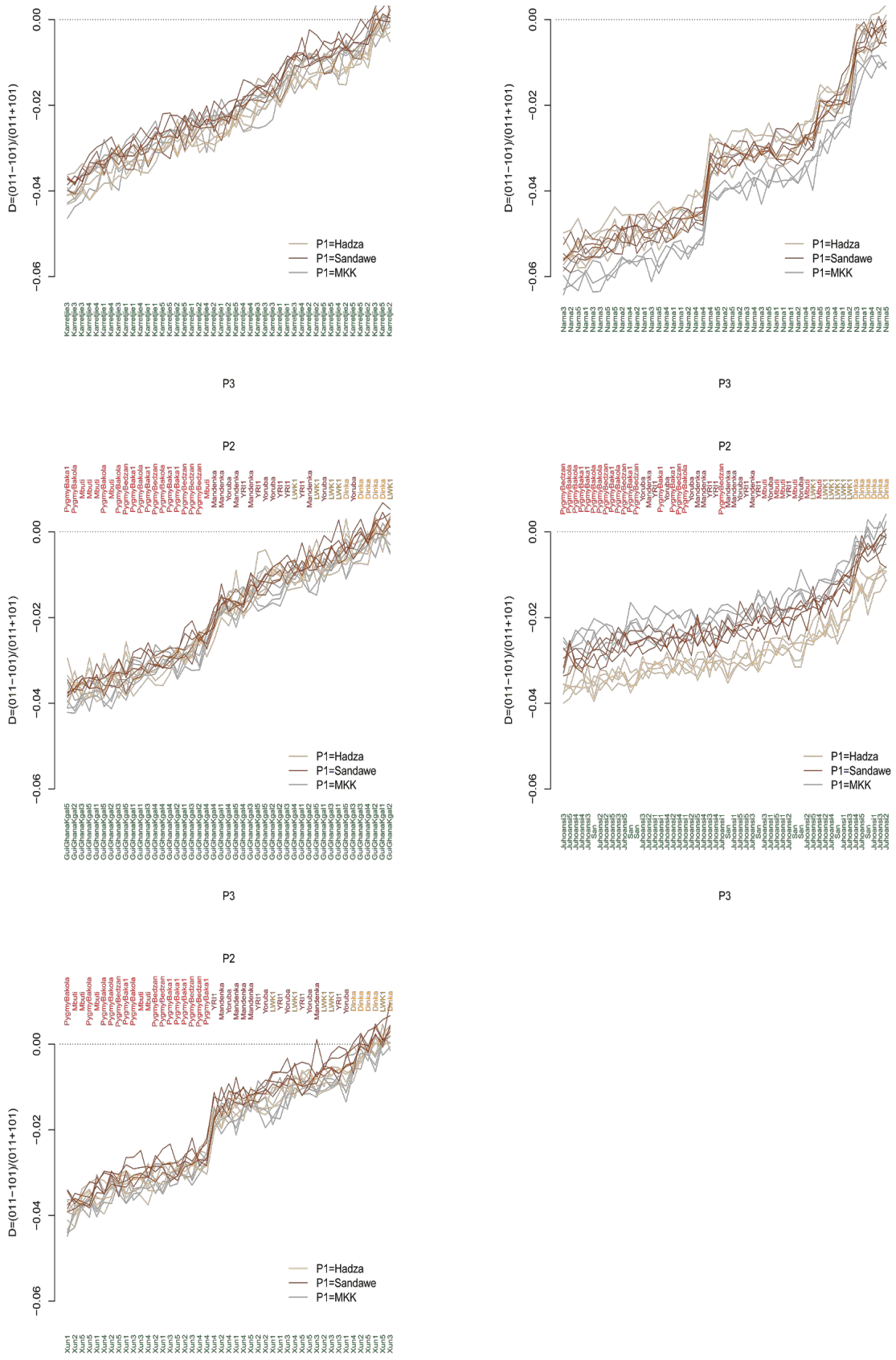


Figure S6.17: D tests Global dataset highlighting eastern African contributions to Khoe-San. Other Africans vs. Khoe-San. Standard deviations not shown.

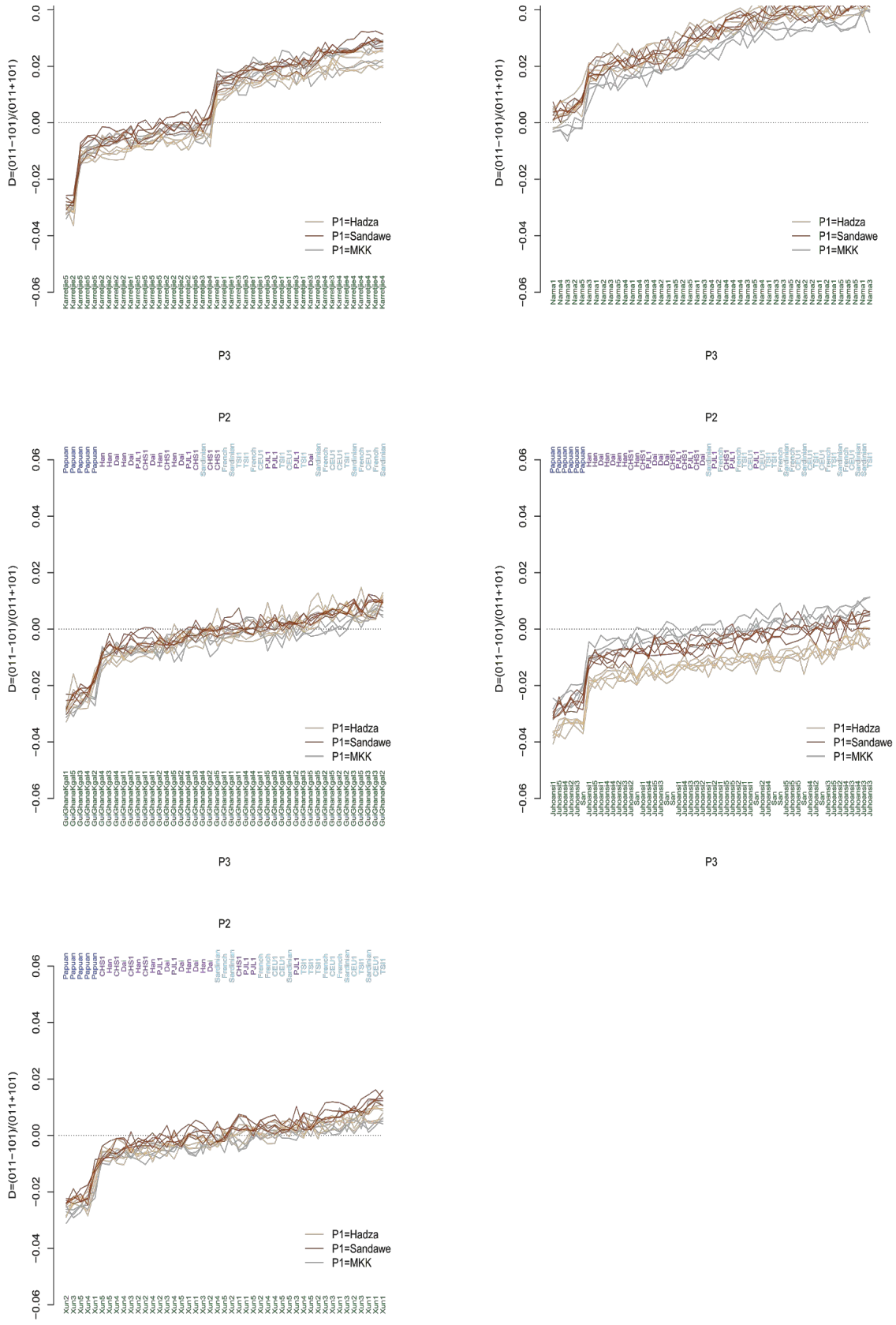


Figure S6.18: D tests Global dataset highlighting eastern African contributions to Khoe-San. Non-Africans vs. Khoe-San. Standard deviations not shown.

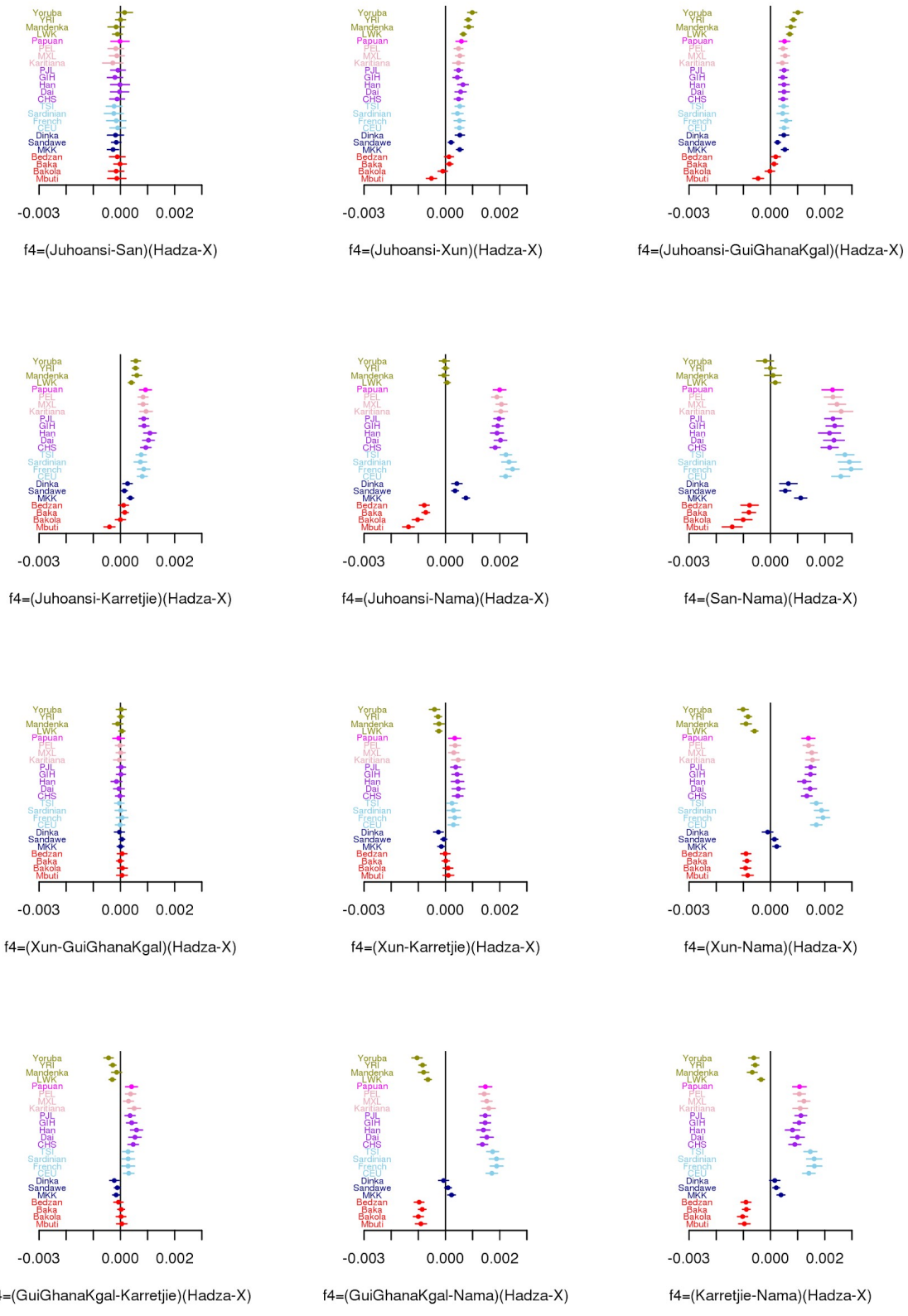


Figure S6.19: f_4=(KSP1-KSP2)(Hadza-nonKSP). A positive value suggests affinity between either KSP1 and Hadza or KSP2 and X while a negative values suggests affinity between either KSP1 and X or KSP2 and Hadza. Assuming that Hadza only has affinity to Ju|'hoansi/San, these patterns reflect several events: 1) a relatively large component from a Bantu speaking source in !Xun and |Gui and ||Ganal, 2) a large European component in Nama, 3) a smaller, but still substantial, non-African component in Karretjie, 4) a smaller Bantu component in Mbuti compared to the other rainforest hunter-gatherer populations (Bedzan, Baka, Ba.Kola).

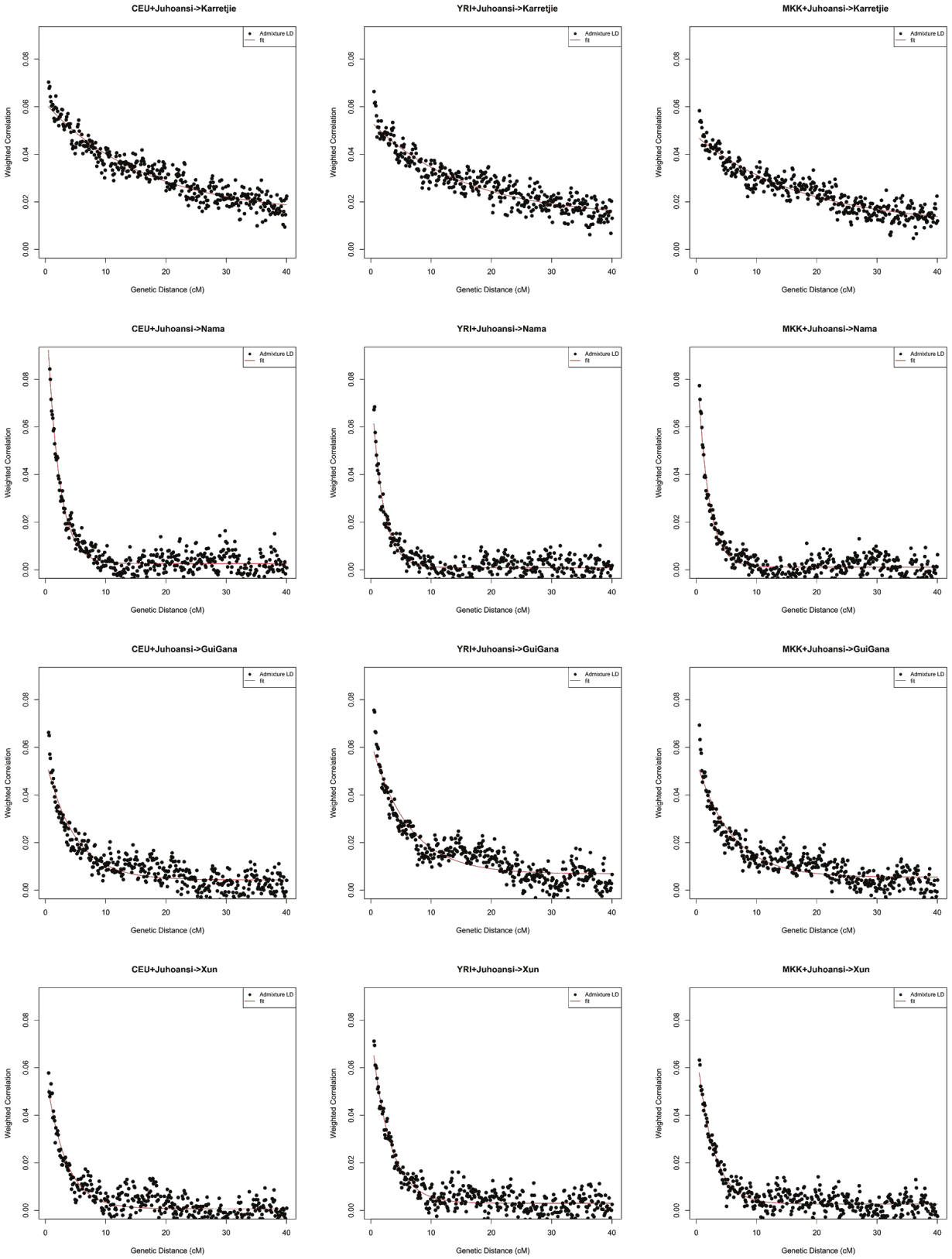


Figure S6.20: LD decay patterns - Admixture of outside groups (CEU YRI MKK) into the various Khoe-San groups.

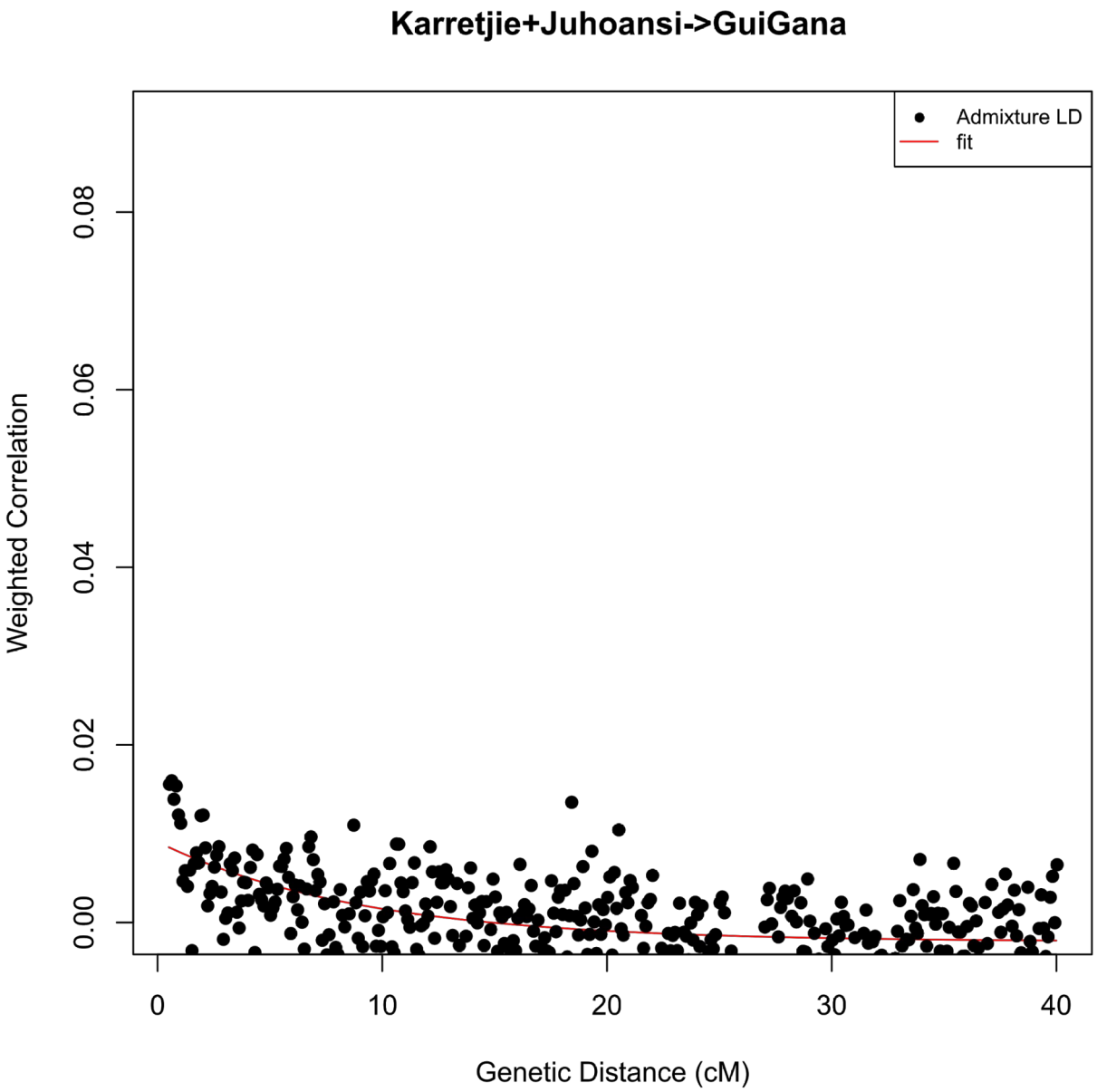


Figure S6.21: LD decay patterns - Central San as admixed group between northern and southern San.

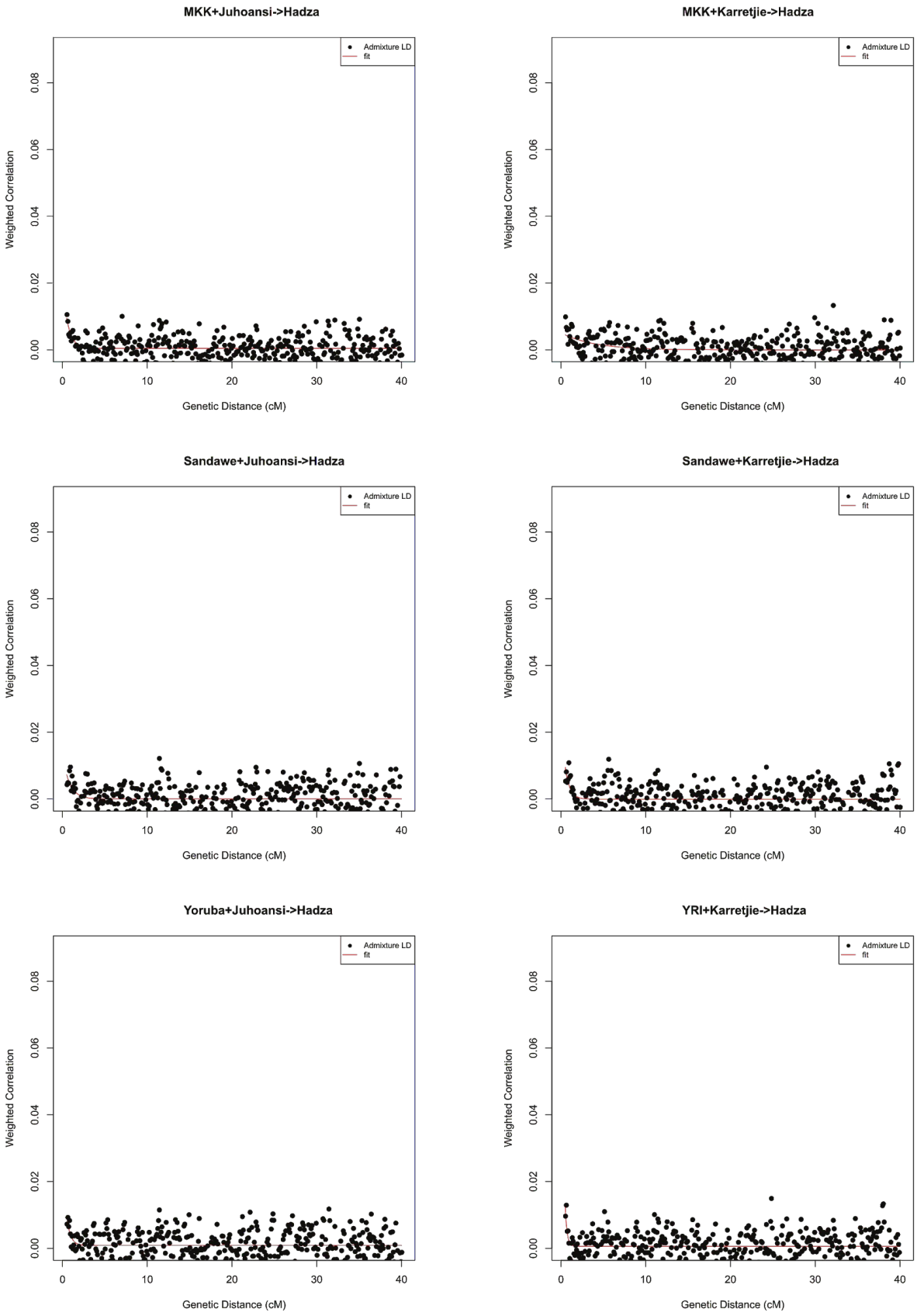


Figure S6.22: Admixture into Hadza with Ju|'hoansi/Karretjie as one source and western/eastern Africans as other source.

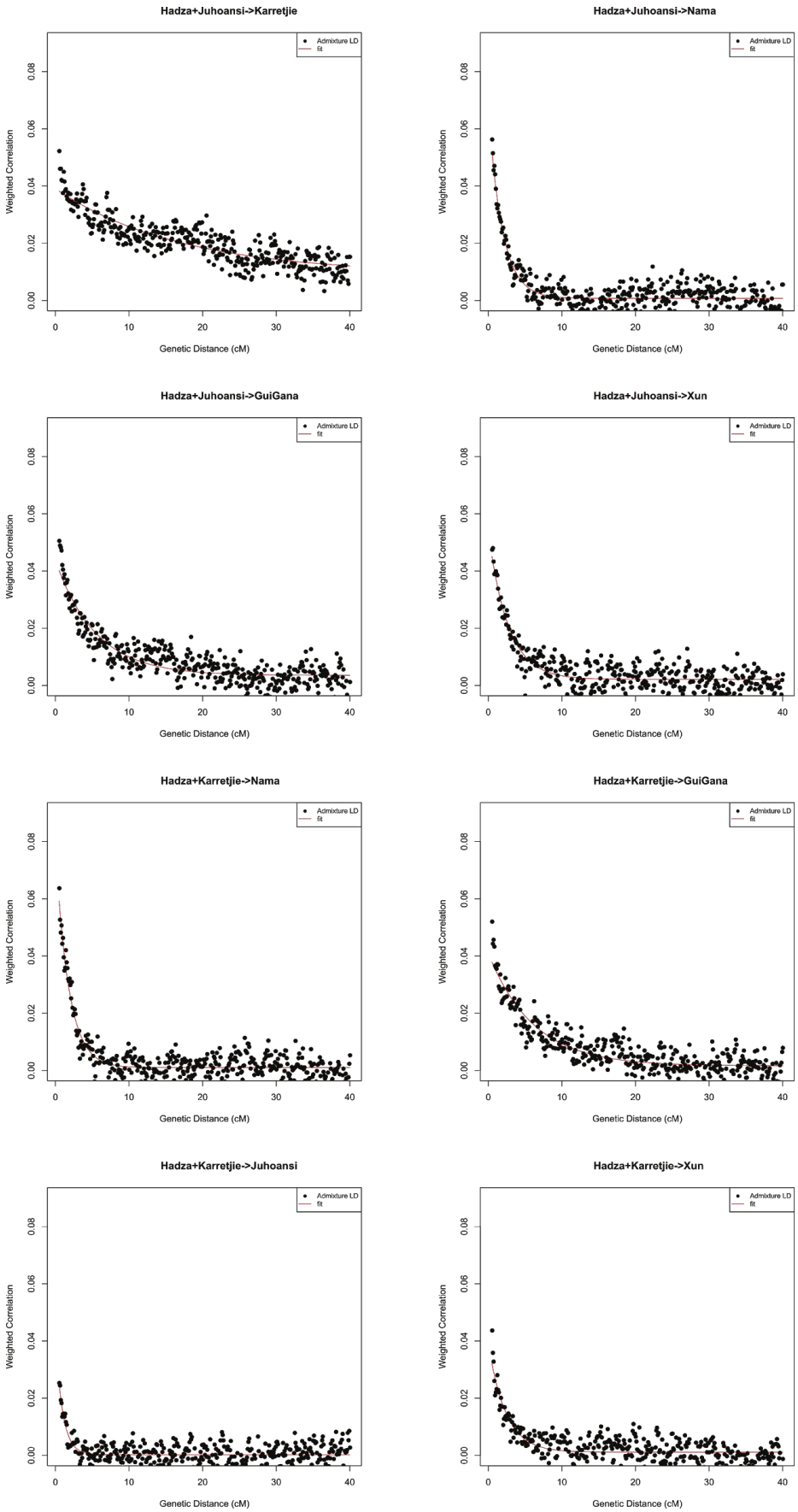


Figure S6.23: LD decay patterns - Hadza and Karretjie/Ju|'hoansi as sources and admixture into various other San groups.

7. Demographic inferences

7.1 Coalescence analysis

7.1.1. Preparation of the dataset for GPhoCS analysis

The sequence data for coalescence analysis was prepared according to the guidelines outlined in (Gronau et al. 2011). Over 30,000 short sequence fragments were sampled from random positions across the autosomes. The length of the fragments was set to 1 kb, which according to (Gronau et al. 2011) is a good length for human genomes, as it represents the optimal trade-off between minimizing the impact of recombination and maximizing information for coalescence analysis. For filtering of the fragments we followed the broad guidelines and recommendations of (Gronau et al. 2011). Five filters were downloaded from the UCSC genome annotation database for hg19 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/), which targets known genic regions (refGene, knownGene), simple and complex repeat regions (simpleRepeat, genomicSuperDups) and CpG islands (cpgIslandExt). In addition we also compiled a filter from our own called indel regions in the Khoe-San dataset. Positions were set to missing using the six different filters and thereafter 1 kb fragments containing more than 10% missing data were filtered out. The pipeline thus contained the following steps; random sampling of 1 kb fragments from the autosomes, marking positions present in filters as missing, filtering out of fragments containing over 10% missing data and, converting the data to the right input format for GPhoCS. This pipeline is then run until over 30,000 fragments were obtained. The exact number of fragments used in the GPhoCS run was 31,622 fragments.

7.1.2. Running GphoCS

GPhoCS was run for all pairwise combinations between all individuals from the HGDP dataset and two individuals each from the five Khoe-San populations. The two individuals (one Nama and one Karretjie) that previously showed data problems were avoided. For the Karretjie, where clear differences between more admixed and less admixed individuals could be seen from other analysis – see Figures S6.1, S6.4, S6.16 –, the two least admixed individuals were chosen. For the rest of the Khoe-San dataset two individuals were randomly chosen.

GPhoCS was run with default input parameters except that the data was logged every 20 steps instead of every 10. No migration bands were added. The MCMC was run for 200,000 iterations and the first 50,000 was discarded as burn in. The visualization of the trace files showed that both the inferred split time (Tau) and population size (Theta) had already stabilized before reaching the burnin cut-off.

7.1.3. Plotting results

Mean and Median split times (Tau) was calculated for the 150,000 remaining logs of Tau after the burn in was removed and are shown in Table S7.1. To convert Tau to years before present a mutation rate of 1.25×10^{-8} and a generation time of 30 years was used.

Mean split times together with standard deviations was further visualized as violin plots (Figure S7.1) and density plots (Figure S7.2) of the pairwise split times was created with ggplot2 in R. The distribution of individual split times of all Khoe-San individuals were also visualized as paired curves (column 1, Figure S7.3).

7.1.4 Checks

To check the repeatability of the GPhoCS run itself, the same dataset of 31,622 fragments was subjected to a second GPhoCS run. The outcome was very similar to the first GPhoCS runs (compare second column of Figure S7.3 to first column).

To check the similarity of outcome if a new set of 30,000 fragments is sampled, a new random sampling of the whole genome was done and 31,691 fragments were subjected to an independent GPhoCS run, using the same input parameters. Although not exactly similar results to the first run

were obtained, the same trends were observed (compare third column of Figure S7.3 to first column).

7.1.5. Discussion

From the coalescent analysis, it is clear that the Khoe-San population represents one of two groups that capture the deepest split of humans. Following the non-Khoe-San leg, rain-forest hunter-gatherers groups diverge, followed by other groups. We dated the mean population divergence of all Khoe-San populations from all other groups to 241 kya (Table S7.1). When only considering the split times of single Khoe-San groups (Table S7.1, Figure S7.2) against other populations, the Ju|'hoansi had the deepest split (266 kya) while the Nama had the most recent split (214 kya). We interpret this could be partially explained by the influence of admixture in the different Khoe-San groups (Figure S7.2, Figure S7.4). The rain-forest hunter-gatherer split was dated to 218 kya and the other splits are subsequent (Table S7.1). We inferred a mean split of 157 kya between the different San groups.

7.2. Inference under a split model with pairwise sampling

Here we follow the approach in (Schlebusch et al. 2017) to estimate model parameters under a pure split model using single individual samples. The assumptions of the model is an infinite number of sites/small mutation rate per site, independence between sites, a pure split model (no migration between branches) and a constant size ancestral population. It does not assume anything about the population size dynamics in any of the populations at times more recent than the split event and neither that the number of generations are the same since the split to the two sampled populations. The parameters that were estimated were T1 (the number of generations from sample 1 to the split multiplied by the per site and per generation mutation rate), T2 (the number of generations from sample 2 to the split multiplied by the per site and per generation mutation rate), α_1 (the probability of not coalescing before the split in branch 1), α_2 (the probability of not coalescing before the split in branch 2) and θ_A (the size of the ancestral population multiplied by the per site and generation mutation rate). The times in years t1 and t2 are then estimated by dividing T1 and T2 by the per site and per generation mutation rate and multiplying by an assumed generation time in years. The ancestral population size is estimated by dividing θ_A by the per site and per generation mutation rate. The expected number of generations to coalescence given that the two lineages coalesce before the split also enter the calculations and we refer to these by V1 (the value for branch 1 multiplied by the mutation rate per site and per generation) and V2 (the value for branch 2 multiplied by the mutation rate per site and per generation). By sampling pairs of genes it is possible to estimate of α_1, α_2, T1, T2, θ_A, V1 and V2 using the counts of the different sample configurations (see (Schlebusch et al. 2017)).

Estimates are obtained by plugging in sample configuration counts into formulas and branch specific drift are given by $-ln(\alpha)$. Effective populations sizes back to the population split can then be obtained by multiplying the estimated drift by the estimated split time.

However, as we will show in the upcoming article, the assumption of a constant ancestral population size is potentially problematic, especially when there is a shared bottleneck just prior to the population split. In this case split time estimates can be negative. The reason is that the split time estimates (using the sample configuration counts) under a model with constant ancestral size are exactly the same as the estimates for

$$T+L(\theta_B-\theta_A)/\theta_B$$

in a model with an initial (scaled) population size θ_B for lasting for mutation scaled time L (number of generations multiplied by the mutation rate) before it changes to the ancestral size θ_A.

We utilized a weighted block jackknife procedure with 5 Mb blocks to estimate the confidence intervals of the parameters. We applied this method to the group-called dataset KSP+HGDP. Careful attention was paid to get the correct number of non-variant sites with the same filtering criteria as variant sites (see above for ancestral calling).

We considered 6 population splits:

1) Khoe-San branch vs non-Khoe-San

2) Northern Khoe-San branch vs southern Khoe-San

3) Rain-forest hunter-gatherer branch vs eastern African+western African+non-African

4) Western African vs eastern African+non-African

5) Non-African vs eastern African branch vs the non-African

6) Western Africa vs western Africa (Mandenka vs Yoruba)

We will refer to these splits as the KSP split, northern KSP split, rainforest hunter-gatherer split, western African split, Out-of-Africa split and the Archaic split. Note that some of these branches are represented by very few individuals (the rainforest hunter-gatherer branch and the eastern African branch consist of a single individual from Mbuti and Dinka respectively).

Results and discussion

The estimates of the ancestral effective population size is remarkably constant across comparisons (Figure S7.5): regardless if a European individual is compared to another European individual or a Khoe-San individual is compared to a western African individual, the ancestral population size is estimated at around 17 thousand individuals. This probably reflects the fact that all humans share a recent past and that (for instance) the out-of-Africa event is a very recent event compared to the average coalescent time of any two lineages.

Split time estimates (Table S7.2, Figure S7.6) are overall very similar to those obtained by GPhoCS but in contrast to GPhoCS, the TT-method estimates negative split times when contrasting non-African individuals - consistent with the out-of-Africa bottleneck ($\theta_B<\theta_A$ in the equation above). The out-of-Africa split (Dinka vs non-African) is consistently estimated at just below 100 kya.

Estimated drifts (Figure S7.7) are consistently lower (suggesting larger effective population sizes) in Khoe-San individuals than in other individuals but as we show below, this is likely to be a result of admixture.

Estimates when contrasting Khoe-San individuals to each other are not performing well (Figure S7.8). Estimates of α are often larger than 1 suggesting model assumptions are not met probably reflecting that a model of isolation by distance would be a better choice of model in order to study the recent demographic history of the Khoe-San.

7.3. PSMC and MSMC

7.3.1. PSMC

We used the Pairwise Sequentially Markovian Coalescent model (PSMC) to estimate effective population size changes over time using a single (diploid) individual (Li and Durbin 2011). Individual input files for PSMC were generated from the “KSP+HGDP” and the “global” datasets VCF files using a in-house converting script. PSMC was run with the following parameters: -N25 -t15 -r5 -p "4+25*2+4+6". Outputs were plotted with psmc_plot.pl. A mutation rate of 1.25e-8 (per base pair per generation) and a generation time of 30 years were used.

The plots for the “KSP+HGDP” dataset are shown in Figures S7.9 and S7.10. The five panels of Figure 7.9 show the curves for each of the five Khoe-San populations; Figure S7.10 shows the curve for all the samples. In general samples from the same population have very similar curves except for the recent times. In the fourth panel (second row right) we can see that the curve of the HGDP San is very similar to the curves of the five KSP Ju|'hoansi. All samples present a decrease in effective population size estimates from ~80 to ~20 kya, with a lowest level at ~40 kya. This decrease is stronger for the non-African populations. These observations are confirmed on Figure S7.11 which shows the curve for all the individuals in the global dataset. In each panel, a different

set of individuals in highlighted (Americans, Europeans, East Asians, Southeast Asians, eastern Africans, western Africans, rainforest hunter-gatherers groups, KS groups). The decrease in population size described above is stronger in the non-African populations, followed by the African non-Khoe-San and finally the Khoe-San individuals. Platform related dataset biases between the samples generated on Illumina vs. Complete Genomics platforms are visible in the different regional representations (Figure S7.11).

7.3.2. MSMC

We used the multiple sequentially Markovian coalescent (MSMC (Schiffels and Durbin 2014)) to estimate effective population sizes over time using one or several diploid genomes. We used the "KSP+HGDP" dataset (both sequenced on Illumina technology) phased with fastPHASE to avoid biases due to the different sequencing platforms. The scripts distributed with MSMC version 0.1.0 (github.com/stschiff/msmc) were used to create input files for the tool. Regions of the genome considered accessible for short paired end reads were obtained from the 1000 genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/20120824_pilot_style_mask.bed). We ran MSMC for single individuals, all possible pairs as well as groups of five individuals per population to obtain insights into different time periods. Default parameters were used except for a fixed recombination rate and a recombination to mutation rate ratio of 0.88. Results were then plotted assuming a mutation rate of 1.25e-8 and a generation time of 30 years.

The results for two chromosomes are qualitatively similar to the PSMC results (Figure S7.12). The results displayed represent the averages of all individuals/pairs of individuals per population. The specific power of MSMC lies in analyzing more than two chromosomes. By analyzing the first coalescent event across all analyzed chromosomes, MSMC can give better resolutions in more recent time periods. This is evident in the more refined dip in effective population size seen between 20 and 60 kya, confirming that Khoe-San groups also had a reduction in population size during this period. This had been seen previously in a single San ancient sample (Schlebusch et al. 2017), which is also represented in Figure S7.12. It will be interesting to see what pattern other African populations will show during this period when multiple high coverage genomes become available and can be analyzed. The simultaneous analysis of ten chromosomes adds additional resolution to the last 10,000 years where all populations increase in size with an intermittent dip between 4,000 and 2,000 years ago.

We observed variations between the MSMC curves obtained using one, two or five individuals, in particular around the time where the "dip" in effective population size is observed. In order to investigate whether such a pattern could be due to a bottleneck, we simulated data under a bottleneck model and ran MSMC on the resulting datasets (see section 9).

To investigate the impact of non-Khoe-San ancestry on the MSMC curves, we also ran MSMC for single individuals on the masked data (see sub-section 4.4) for each of the 25 Khoe-San samples as well as for the HGDP San sample. Figure S7.13 shows the MSMC curves for all samples; we plotted the curve based on the entire genome as well as for the Khoe-San homozygous regions ("KSP masked") and the regions which are not homozygous Khoe-San ("KSP masked negative") as a control. The general trajectories are similar between the full data and "KSP masked" regions indicating no strong impact of other ancestries. Most "KSP masked" curves suggest slightly higher effective population sizes during the bottleneck period between 20 and 50 kya, and lower effective population sizes in recent times. The negative mask shows a great variety of different patterns between samples from curves similar to the original pattern to large deviation. This is consistent with the large heterogeneity in other ancestries between individuals and populations.

Demographic inferences Tables and Figures

Table S7.1: Split times inferred by GphoCS.

	Mean Splittime (Tau)	Sdev Splittime (Tau)	Mean Splittime (Generations)	Mean Splittime (Years)	Median Splittime (Generations)	Median Splittime (Years)
Khoe-San vs. Other Pops	1.006	0.12	8047.7	241431.3	8022.3	240669.6
Mbuti vs. Oth (excl Khoe-San)	0.908	0.047	7265.8	217974.3	7295.4	218863.2
Khoe-San vs. Khoe-San	0.652	0.095	5218.2	156547.2	5355.1	160653.6
Western Afr. vs. Oth (excl Khoe-San, Mbuti)	0.524	0.065	4194.1	125821.6	4248.2	127447.2
Eastern Afr vs. Non-Afr	0.46	0.046	3679.6	110387.6	3627.5	108825.6
Non-Afr vs. Non-Afr	0.036	0.016	285.3	8558.7	268	8040
Separating various Khoe-San populations from Khoe-San group as a whole when inferring split times						
Karretjie	1.011	0.11	8091.9	242758.2	7946.2	238384.8
Khoe-San (excl Kar) vs. Oth	1.005	0.122	8037.9	241136.5	8039.4	241180.8
Nama	0.893	0.132	7146.1	214384.1	6814	204420
Khoe-San (excl Nam) vs. Oth	1.031	0.102	8248.1	247441.8	8149.6	244488
|Gui and ||Gana	0.983	0.074	7862.8	235883.4	7847.4	235420.8
Khoe-San (excl |Gui and || Gana) vs. Oth	1.011	0.128	8088.8	242664.2	8099.8	242995.2
Ju|'hoansi	1.108	0.085	8861	265830.1	8910.9	267326.4
Khoe-San (excl Ju|'hoansi) vs. Oth	0.968	0.109	7742.7	232281.8	7776.2	233284.8
!Xun	0.984	0.069	7870.1	236101.5	7822.4	234672
Khoe-San (excl !Xun) vs. Oth	1.011	0.128	8087.2	242615.8	8129	243868.8

Table S7.2: Split times inferred by the TT method.

	Mean Splittime	Sdev Splittime
Khoe-San vs. Other Pops	240,788.5	25,301.12
Mbuti vs. Oth (excl Khoe-San)	214,541.5	8,720.397
Khoe-San vs. Khoe-San	193,619.6	17,672.12
Western Afr. vs. Oth (excl Khoe-San, Mbuti)	117,211.4	5,711.047
Eastern Afr vs. Non-Afr	83,171.02	4,851.04
Ju|'hoansi	264,341.4	12,078.21
Karretjie	242,515.2	22,597.4
||Gui and ||Gana	238,792.1	14,530.04
!Xun	236,538.5	13,382.18
Nama	211,540.4	30,367.15

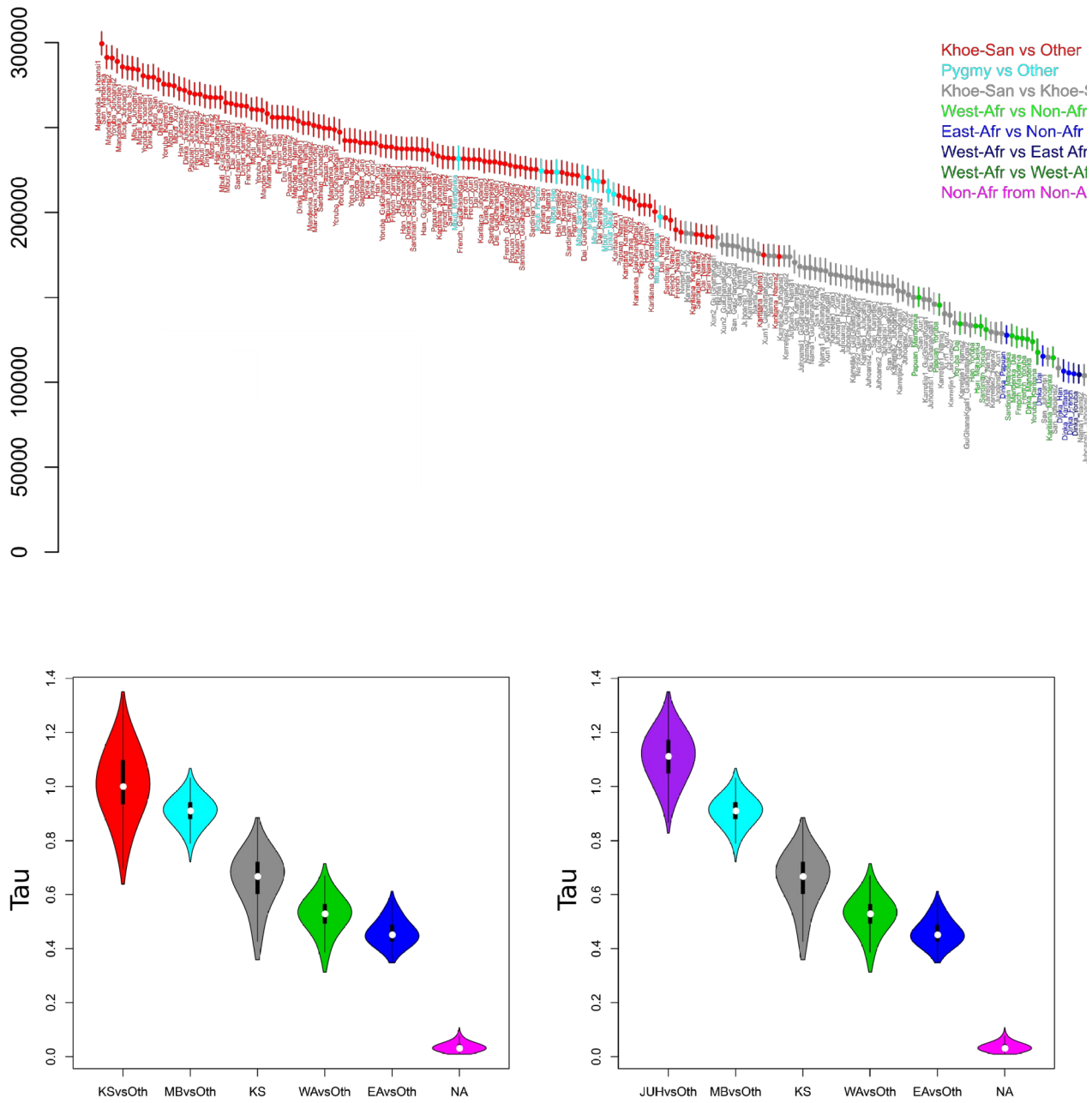


Figure S7.1: GPhoCS analysis. Top panel – Split times sorted descendingly. Colors according to grouped split times indicated in legend. Time was calculated from the GPhoCS inferred unscaled split time (Tau) using a mutation rate of 1.25x10-8 and a generation time of 30 years. Bottom left - Violin plots of grouped split times. Bottom Right - Violin plots of grouped split times where only Ju|'hoansi is used as Khoe-San group.

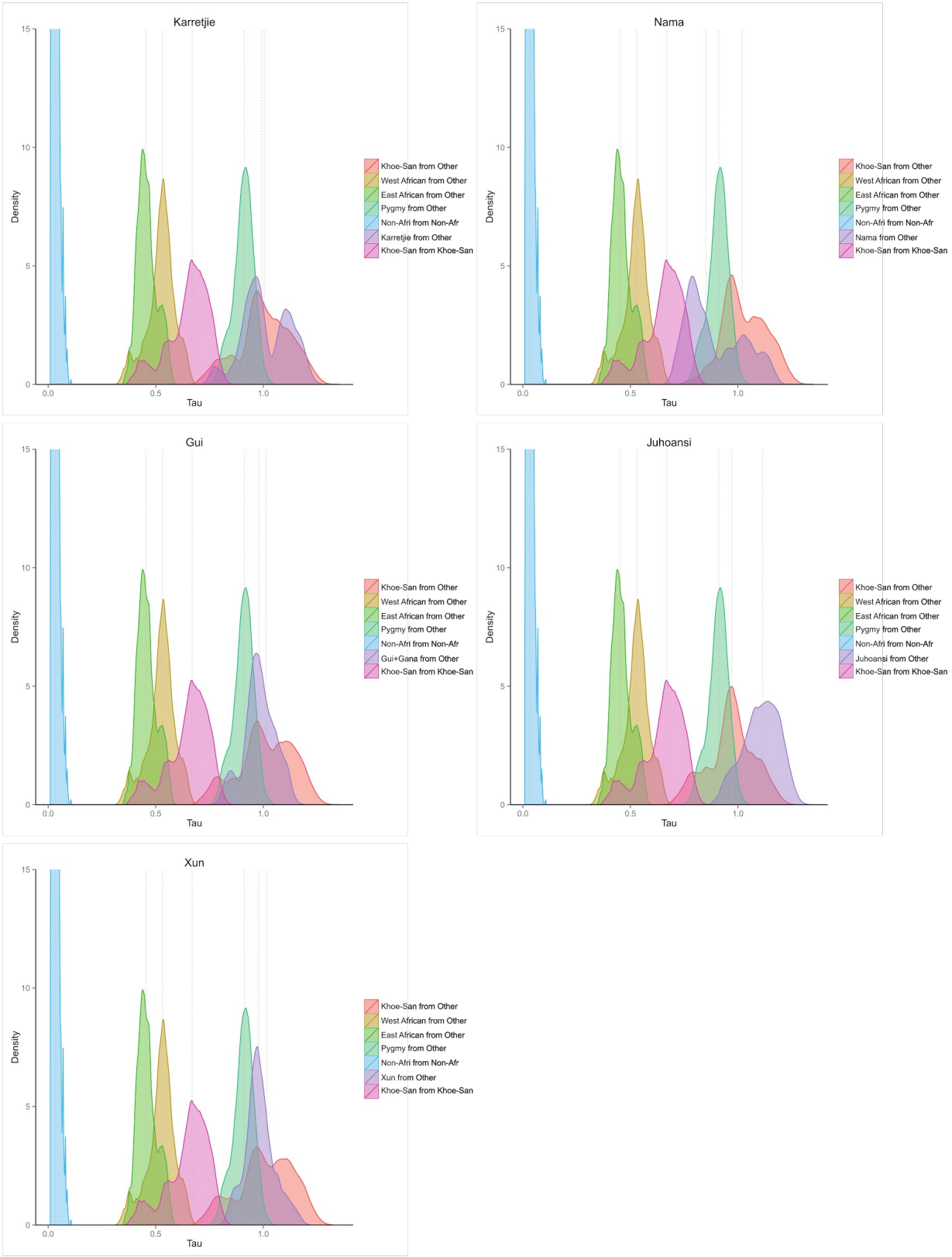


Figure S7.2: GPhoCS analysis - kernel density plots of split times (Tau) of the pairwise comparisons of all individuals in the KSP-HGDP dataset. This figure show the density plots for five instances where one Khoe-San population is plotted separate from other Khoe-San populations. See Main Figure 2 for all Khoe-San groups combined into one group.

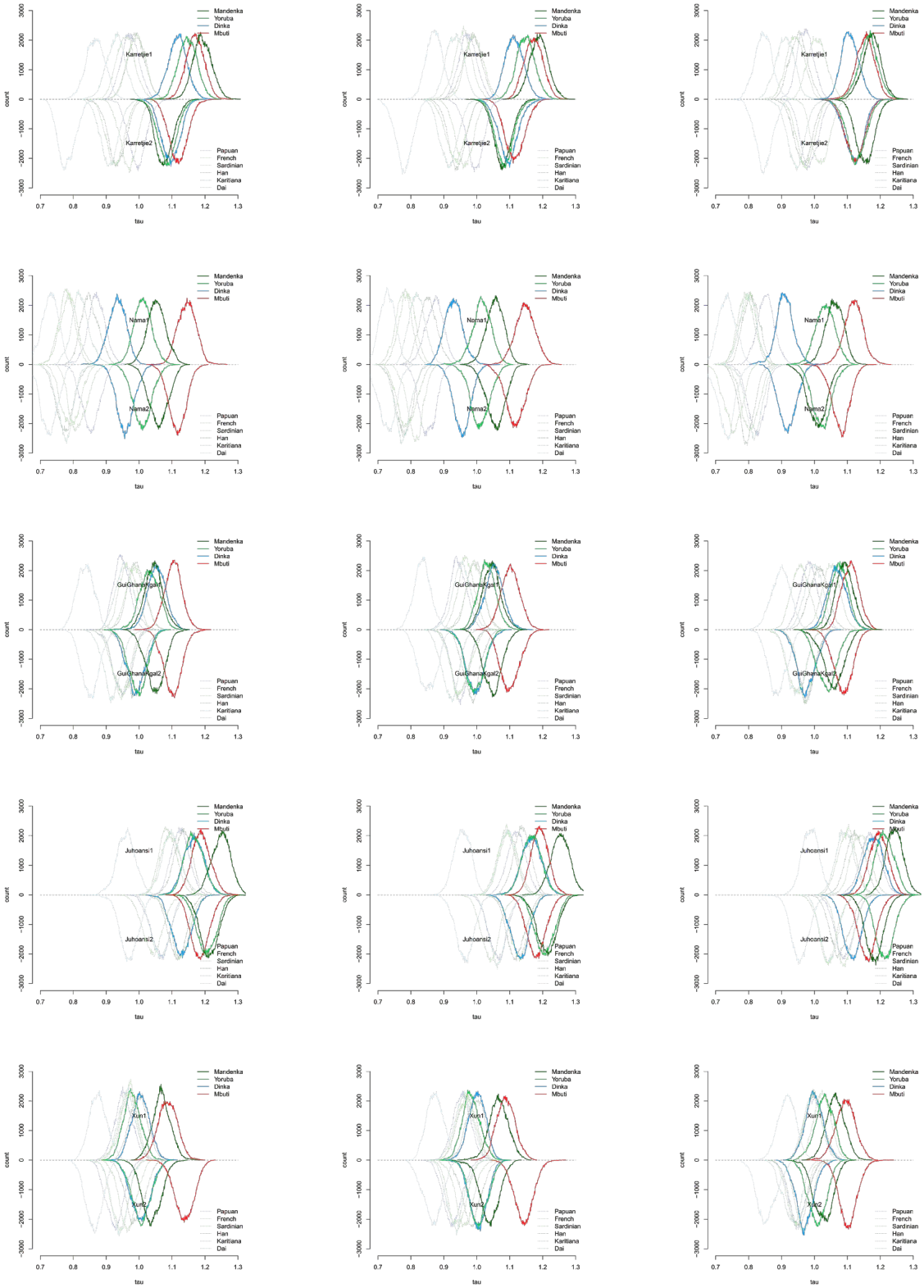


Figure S7.3: First Column Run1 with 31,6252 independently sampled fragments across the genome, Second Column - Run2 - Sample dataset as in Run 1 but independent (repeat) GPhoCS run, Third Column - Run3 - newly sampled dataset of 31,691 independently sampled 1 kb fragments across the genome.

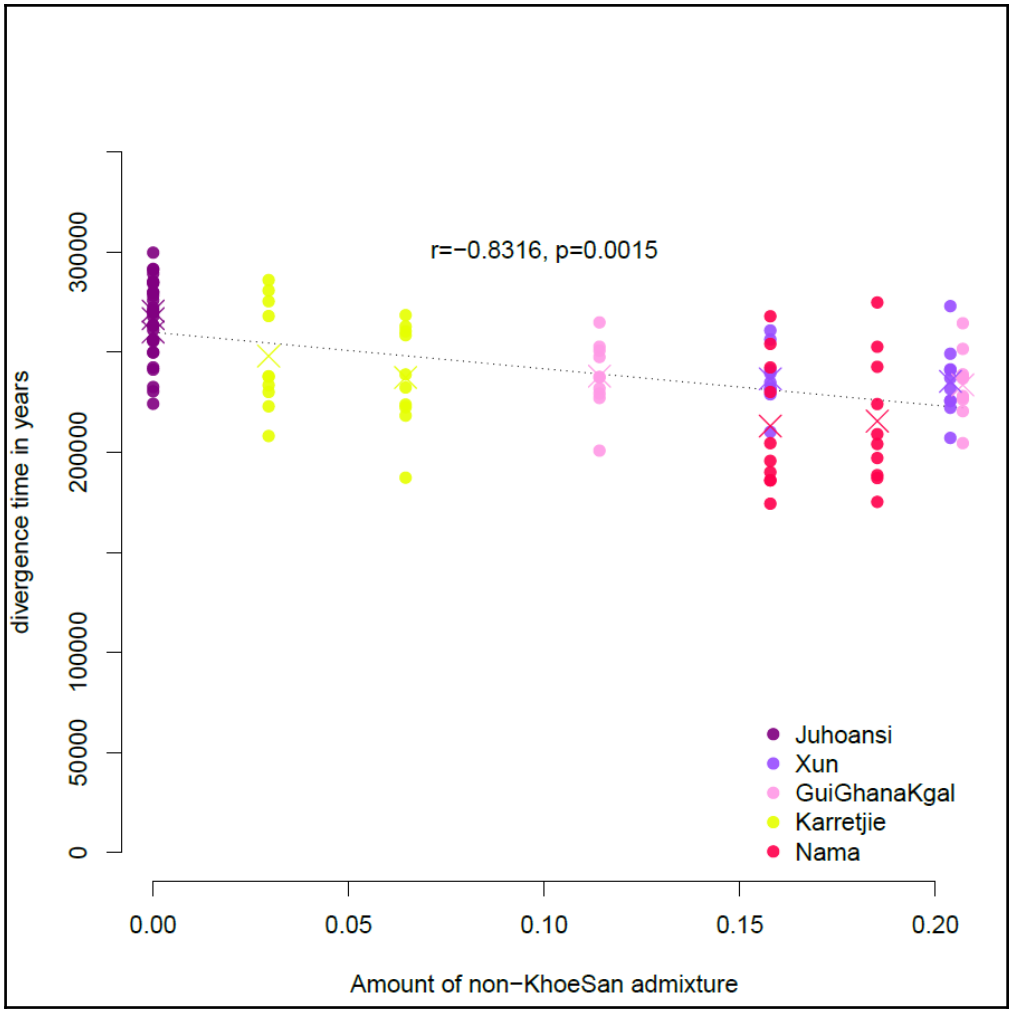


Figure S7.4: Correlation of admixture in Khoe-San vs. divergence times. The amount of non-Khoe San admixture (from the admixture analysis with K=3) on the x-axis and the divergence time based on GphoCS on the y-axis. Three Ju|'hoansi individuals and two individuals from each of the other Khoe-San population were included in the GphoCS analysis. The divergence time with each of these 11 Khoe-San individuals compared to the nine non-Khoe-San individuals (Karitiana, Dai, Sardinian, French, Papuan, Han, Dinka, Yoruba, Mandenka, Mbuti) are depicted with circles. Crosses show the mean across the nine non-Khoe-San divergence times for each Khoe-San individual. The dotted line is the regression line according to Pearson's product-moment correlation.

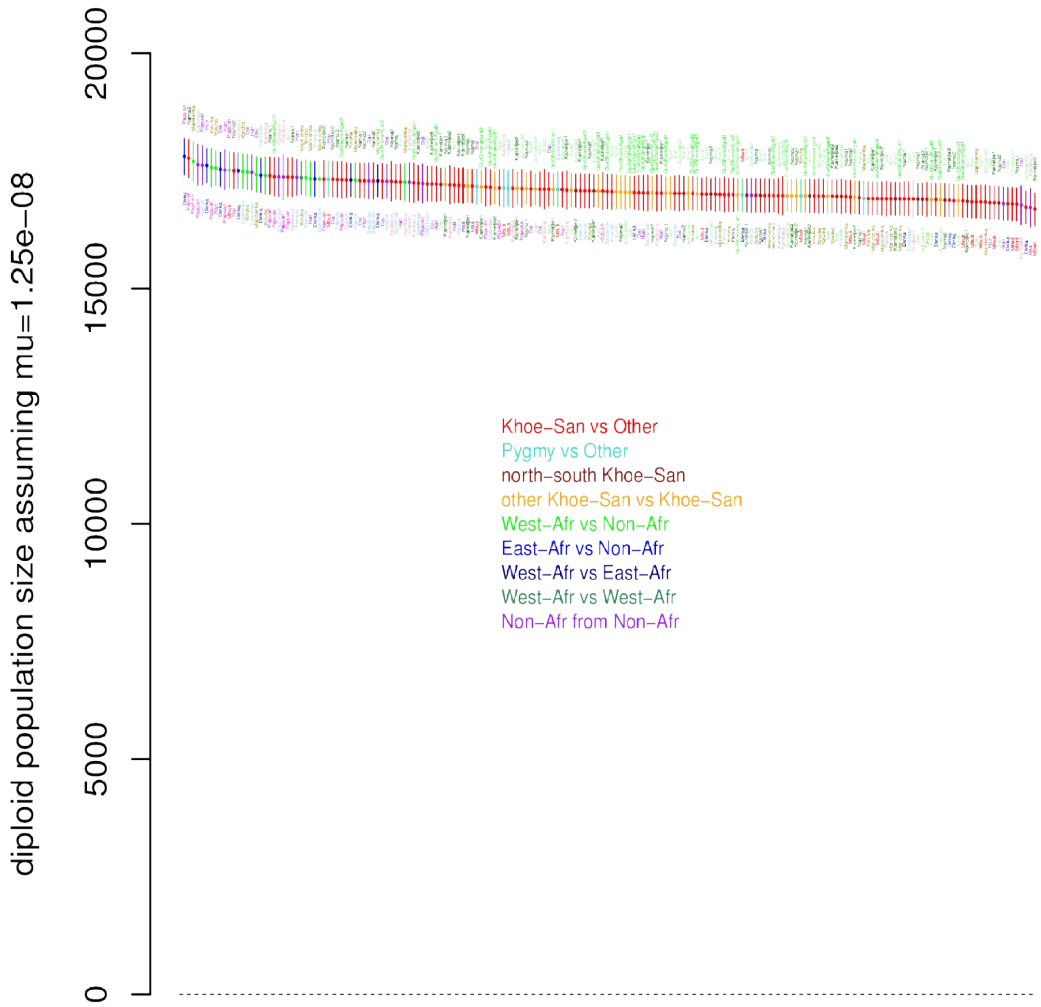


Figure S7.5: Effective sizes of the (constant) ancestral population using the TT-method. The mean ± 2 sd of the estimate of the constant effective size of the ancestral population in a split model. Different bar colors correspond to different splits and the colors of the individuals correspond to different phylogenetic positions.

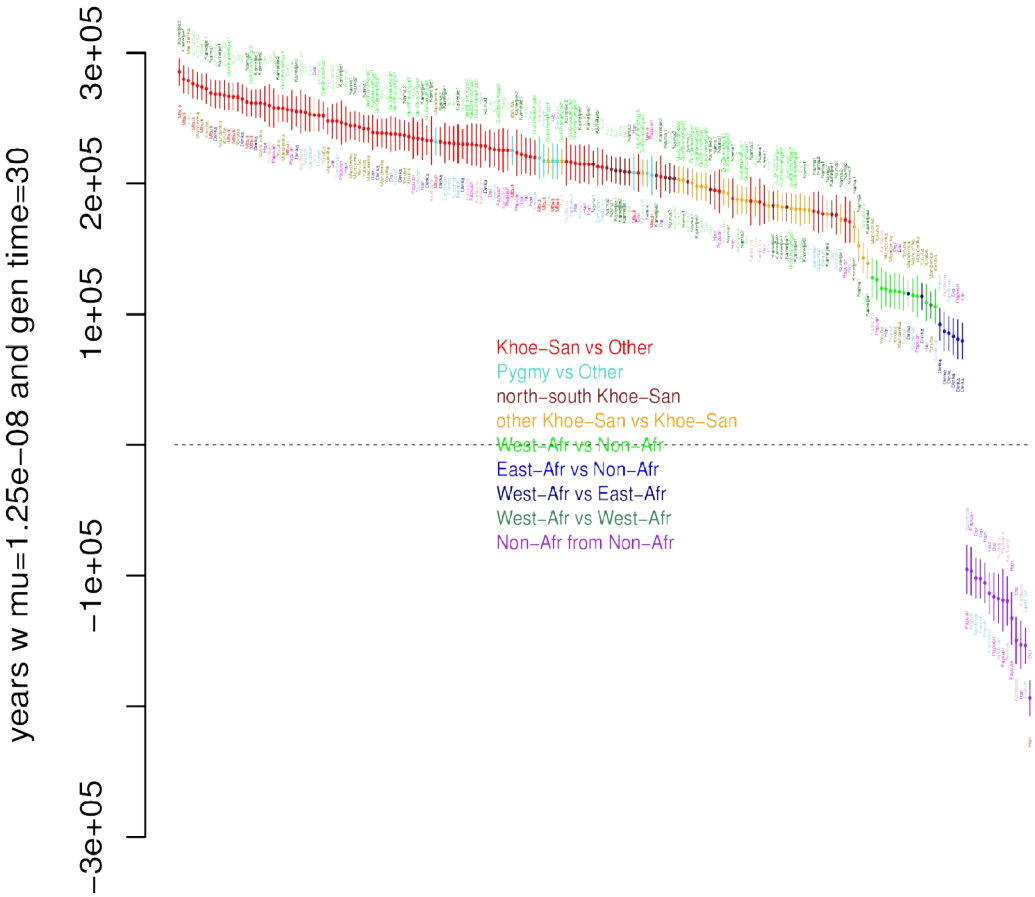


Figure S7.6: Estimated times to split in a split model using the TT-method. The mean ± 2 sd of the divergence time in a split model. Different bar colors correspond to different splits and the colors of the individuals correspond to different phylogenetic positions. The negative estimates of split times outside Africa can be proven to be related to the out-of-Africa bottleneck predating the split in these comparisons.

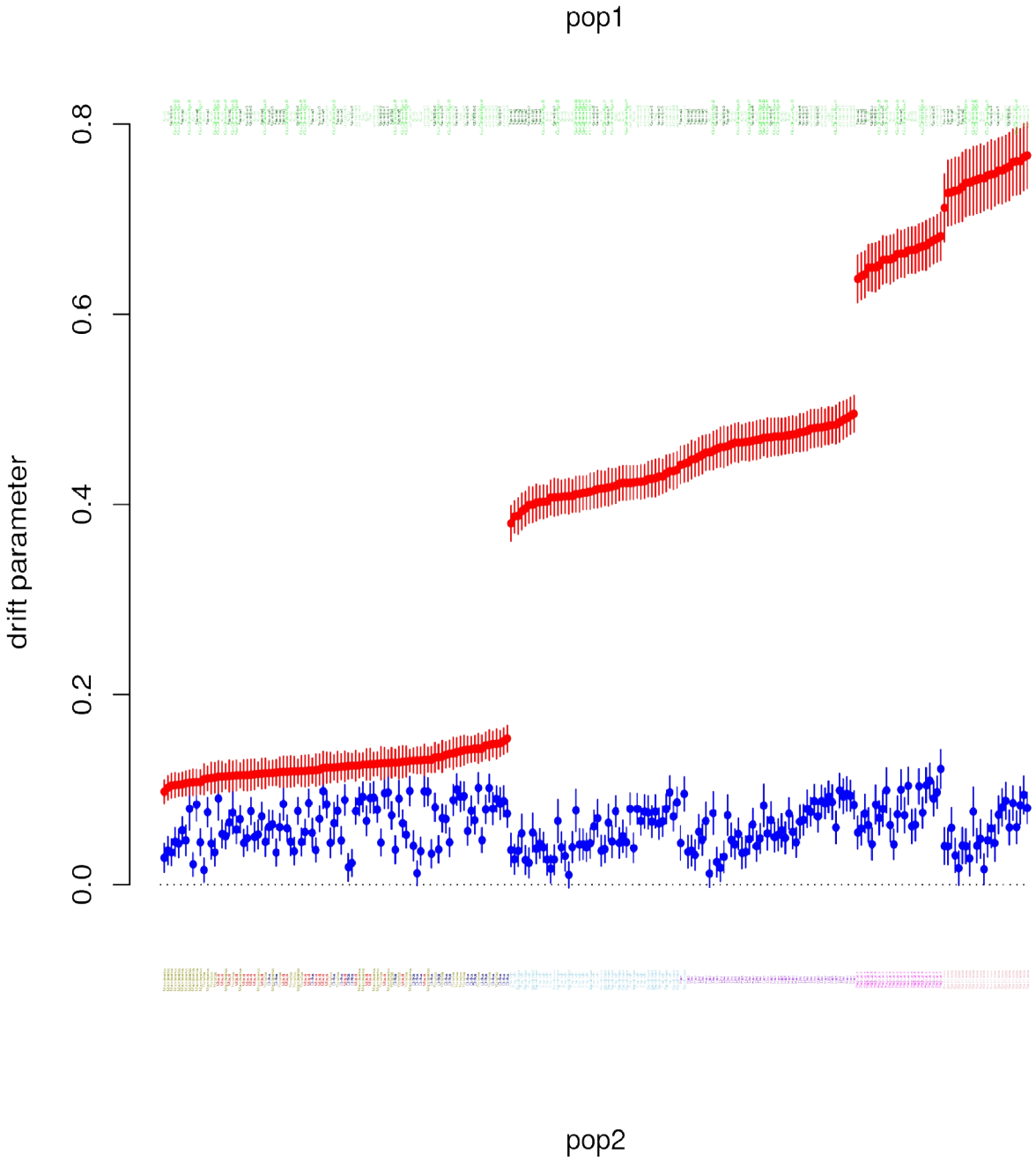


Figure S7.7: Estimated branch drifts using the TT-method since the Khoe-San split. The mean ± 2 sd of drift in a split model. Branch drifts on the Khoe-San branch are shown in blue while branch drifts on the non-Khoe-San branch are shown in red. The name of the Khoe-San individual in each comparison is shown on top and the name of the non Khoe-San individual at the bottom of the figure. Text is brown for western Africans, Mbuti is red, Dinka is blue, European light blue, Han and Dai are purple, Papuan is pink and Karitiana is light pink. Text is light green for northern Khoe-San individuals, dark green for southern Khoe-San individuals and medium green for |Gui and || Gana.

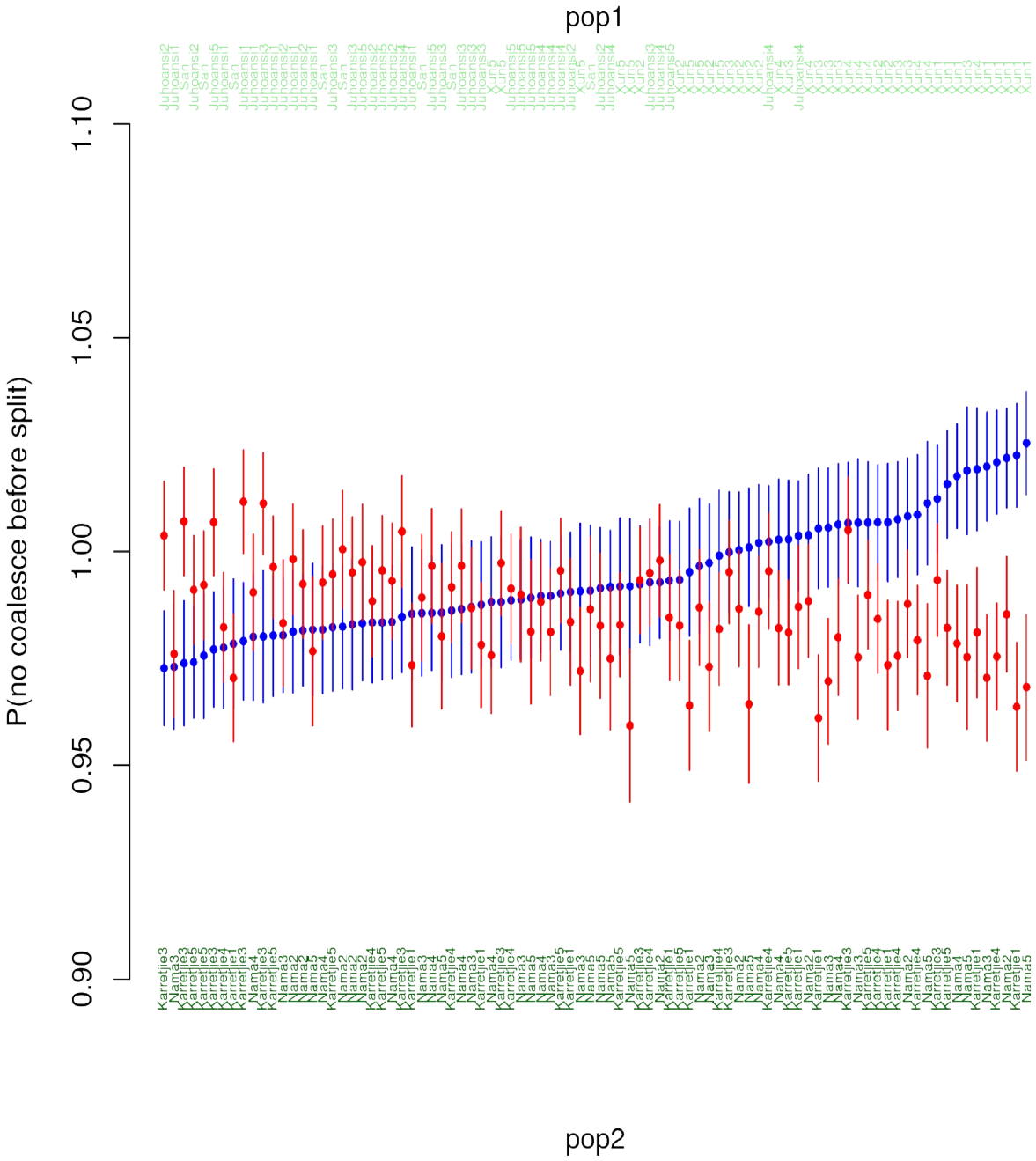


Figure S7.8: Estimated probabilities not to coalesce before the split using the TT-method assuming a split model between northern and southern Khoe-San. The mean ± 2 sd of α assuming a split model. Values on the northern Khoe-San branch shown in blue, values on the southern Khoe-San branch shown in red. The northern Khoe-San individual in each comparison is shown on top and the southern Khoe-San individual at the bottom of the figure.

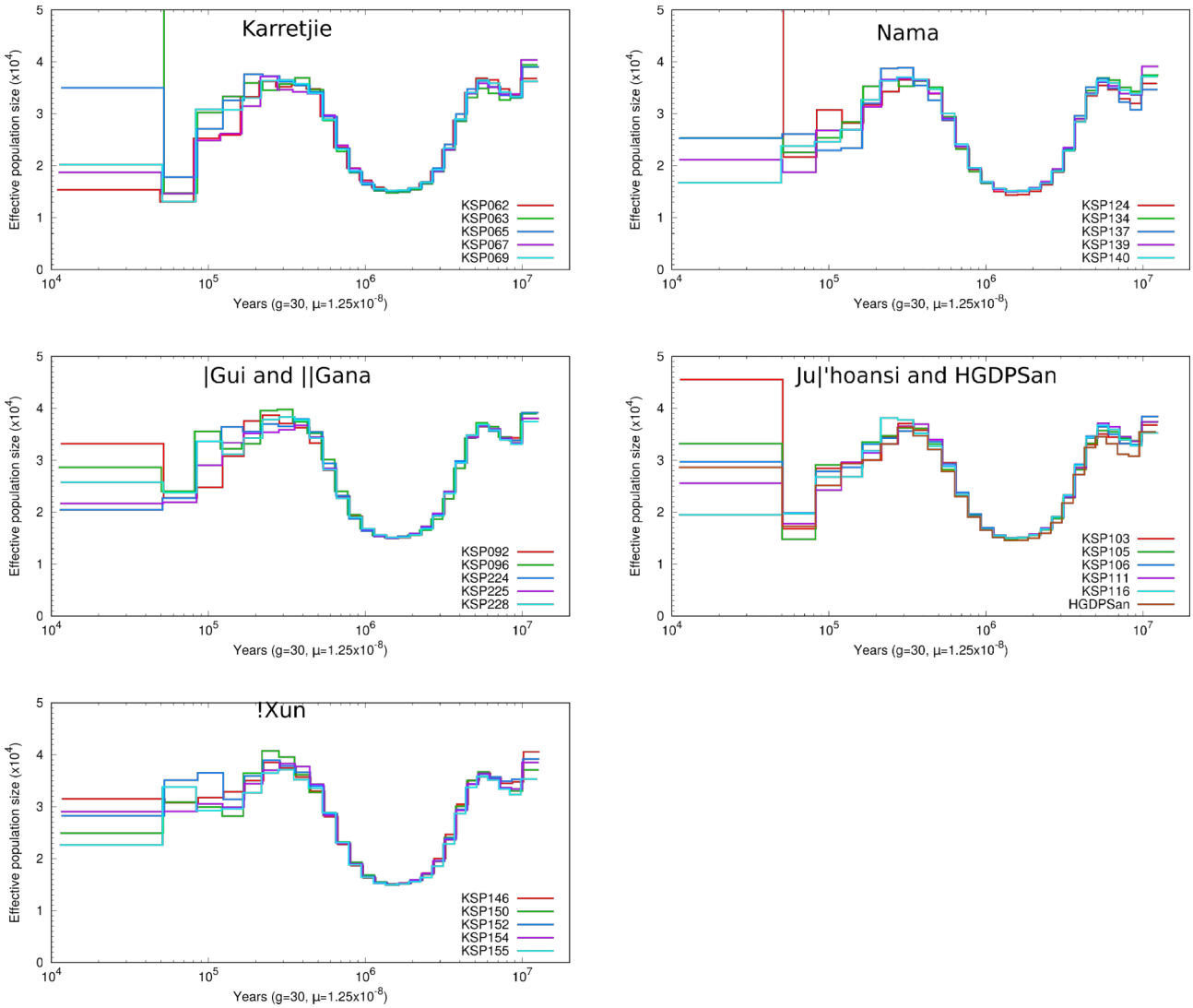


Figure S7.9: PSMC plots KSP dataset and HGDP San. One plot per population. Mutation rate: 1.25e10-8. Generation time: 30 years. x-axis: log.

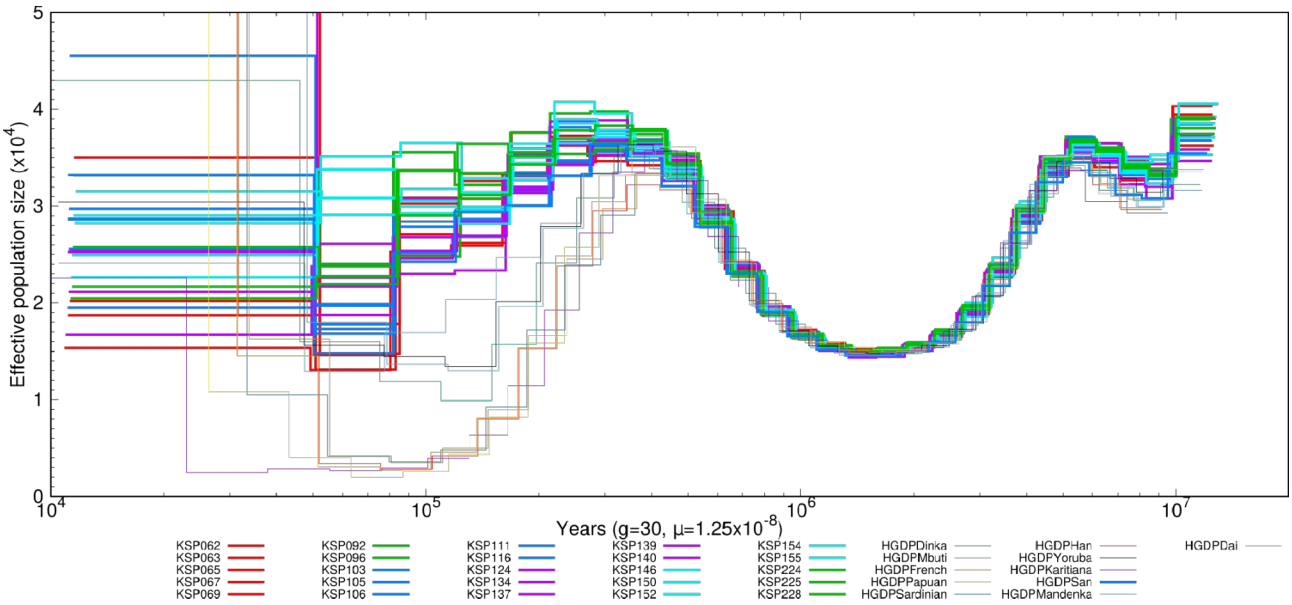


Figure S7.10: PSMC plot KSP dataset and all HGDP samples. Mutation rate: 1.25e10-8. Generation time: 30 years. x-axis: log.

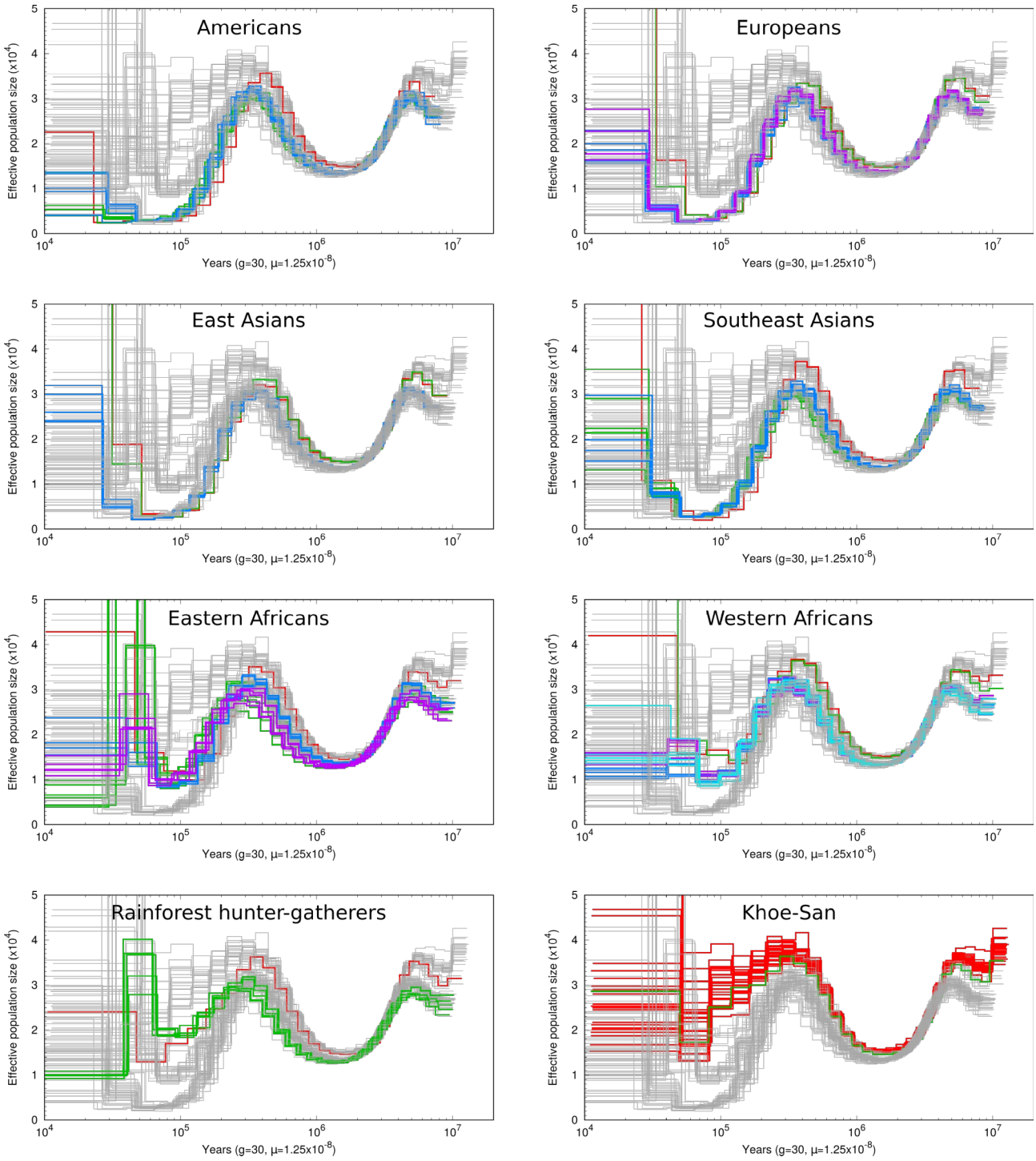


Figure S7.11: PSMC plots of global dataset. Mutation rate: 1.25e10-8. Generation time: 30 years. x-axis: log. Slight shifts in PSMC curves are visible for HGDP (Illumina platform) samples and the CG, LC, and KGP samples (Complete Genomics platform).

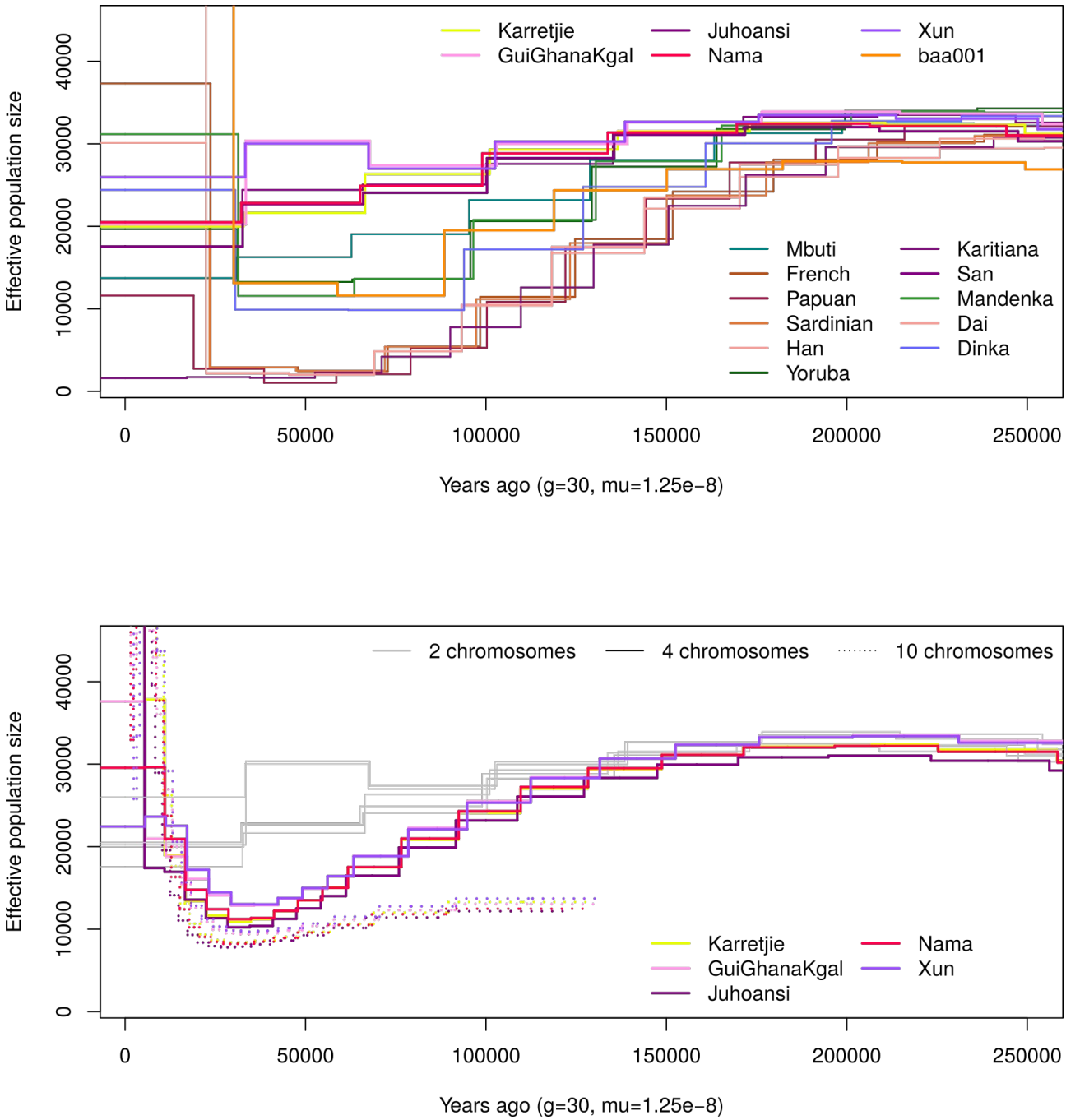


Figure S7.12: MSMC plots. Upper panel: effective population sizes estimated for single individuals (i.e. two chromosomes) for the Khoe-San samples (average over the five individuals in each population), the HGDP samples and the ancient San sample Ballito Bay A ("baa001"). The result for Ballito Bay A has been previously reported (Schlebusch et al. 2017). Lower panel: Khoe-San effective population sizes estimated from single individuals ("two chromosomes", gray), pairs of individuals ("four chromosomes", plain lines) and five individuals ("ten chromosomes", dotted lines). The curves are averaged over all MSMC runs for all different combinations of individuals (respectively five, ten and one).

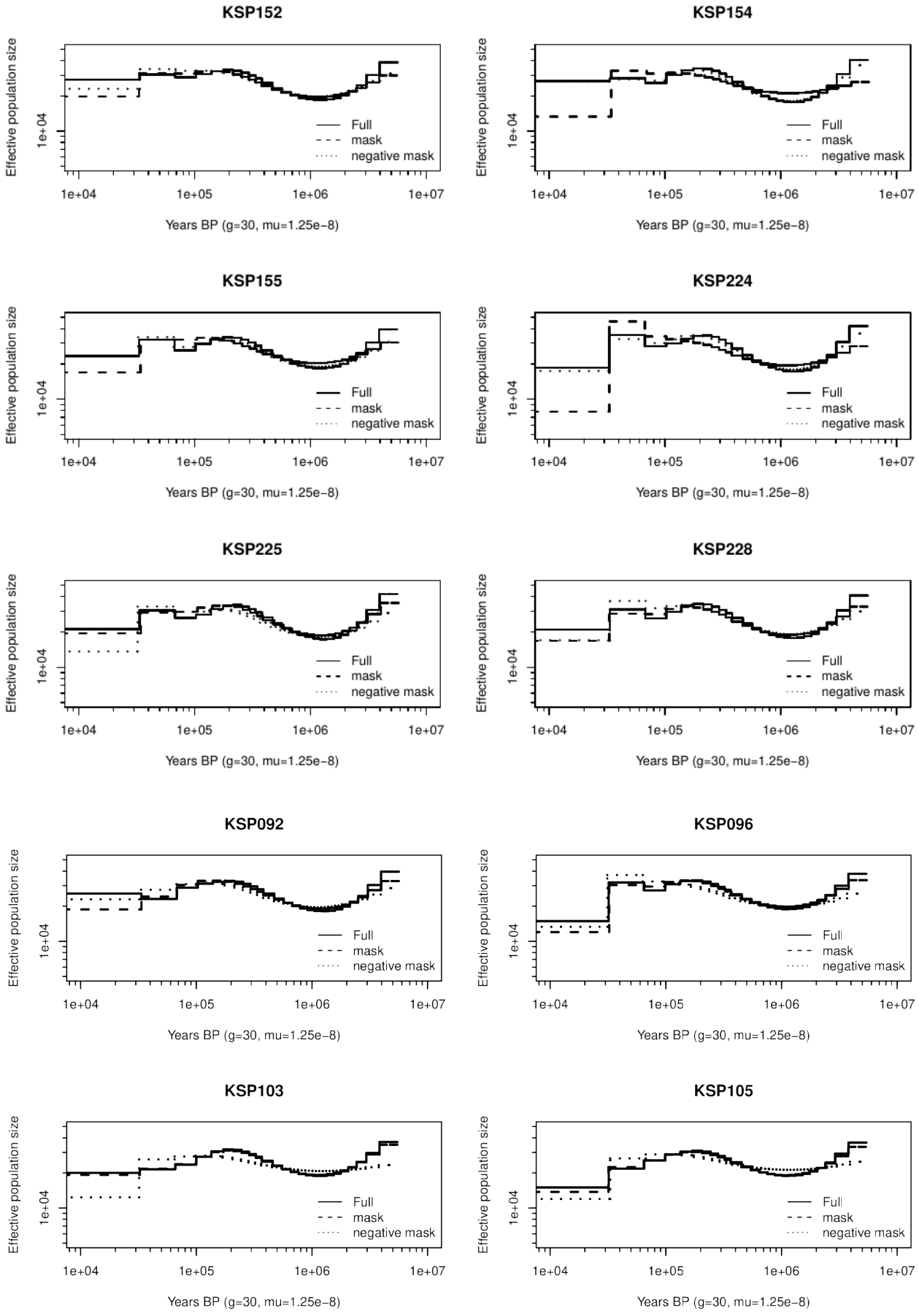


Figure S7.13 (start)

Figure S7.13 (cont.): MSMC results for all Khoe-San individuals separately. Solid lines show the results for the full genome, dashed lines only for regions that were homozygous for Khoe-San ancestry and dotted lines are only regions not homozygous for Khoe-San ancestry.

8. Demographic inferences with Khoe-San specific regions

Heterozygosity was also calculated after masking for the “Khoe-San” portion of each Khoe-San genome as well as for the segments found in more than ten (respectively twenty) Khoe-San-specific masks in all Khoe-San and HGDP genomes. Positions that were masked as “Khoe-San” in more than ten Khoe-San individuals corresponded to about 96% of the sites while positions that were masked in more than 20 Khoe-San individuals corresponded to only approximately 1.3% of the sites. The results are plotted in Figure S8.1.

For the 25 Khoe-San individuals, the heterozygosity decreased when considering only the Khoe-San specific regions. Before admixture masking, the heterozygosities among Khoe-San individuals is significantly higher than among Mandenka, Yoruba and Mbuti (Wilcoxon rank test p-value = 0.0005473). The same test on the heterozygosities for the Khoe-San specific regions has a p-value of 0.01259. The heterozygosities based on regions were more than ten (or 20) Khoe-San individuals were masked was performed in order to check that the lower heterozygosities after masking was not due to a bias in the masking. For instance, if regions with low heterozygosities/mutation rates were more likely to be masked as Khoe-San specific this would result in lower heterozygosities in all individuals -- also the non Khoe-San individuals. Already the fact that very few sites had more than 20 masked Khoe-San individuals suggests that this is not a likely explanation and the results show that, if anything, the heterozygosities increase for such sites.

We also re-estimated split times and branch specific effective population sizes and compared to the original estimates. Compared to estimates based on un-masked genomes, estimates of the split time between Khoe-San individuals and other individuals are consistently deeper and considerably more consistent across Khoe-San individuals when restricting the analysis to the Khoe-San specific portions of the genomes (Figure S8.2). Similarly, estimates of branch specific drifts back to the Khoe-San split are more consistent across African individuals when restricting the analysis to Khoe-San specific regions compared to the non-masked estimates (Figure S8.3 and Figure S8.4).

Demographic inferences with Khoe-San specific regions Tables and Figures

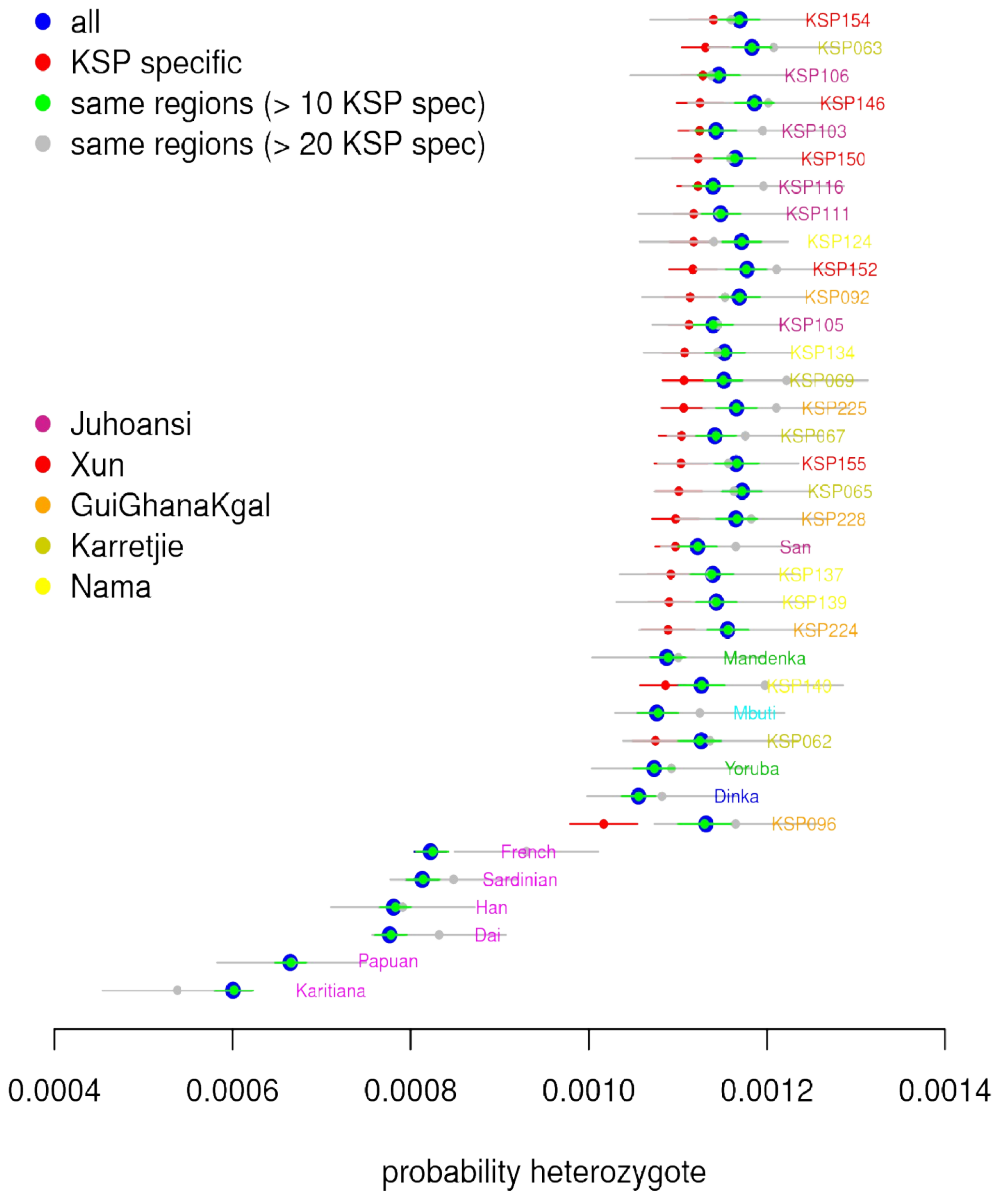


Figure S8.1: Observed heterozygosity per individual using all data (blue), restricted to KSP specific genomic regions (red), restricted to sites where more than ten KSP individuals were KSP specific at this site (green) and restricted to sites where more than 20 KSP individuals were KSP specific at this site (gray).

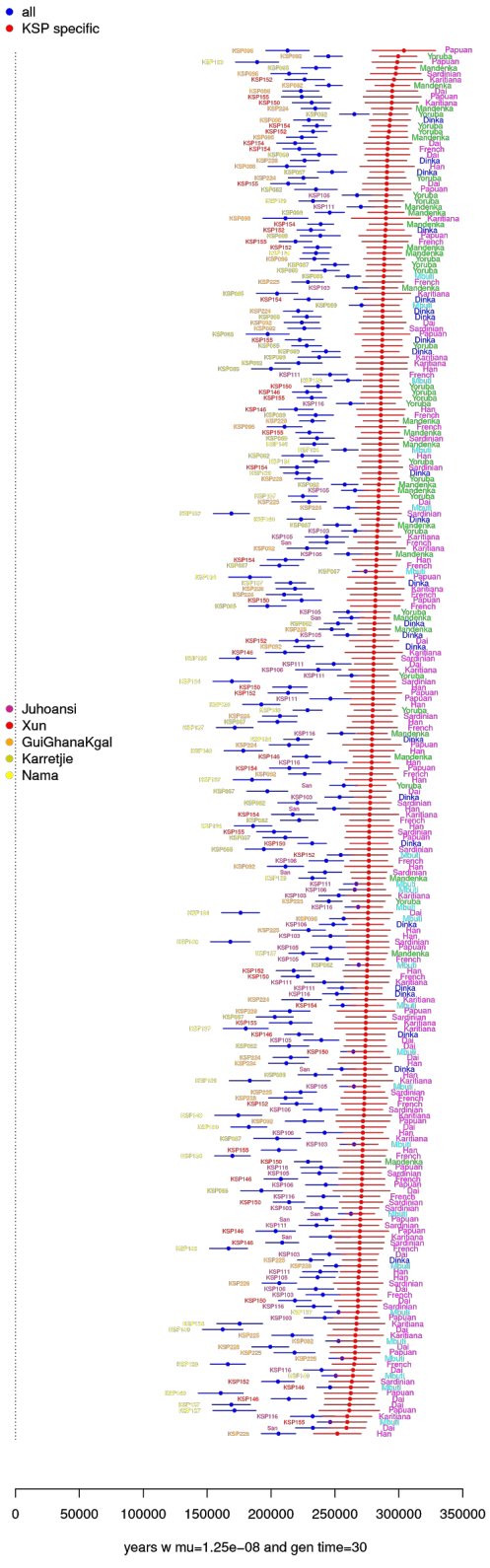


Figure S8.2: Split time estimates with the TT method using all data (blue) compared to when restricting the analysis to KSP specific genomic regions (red).

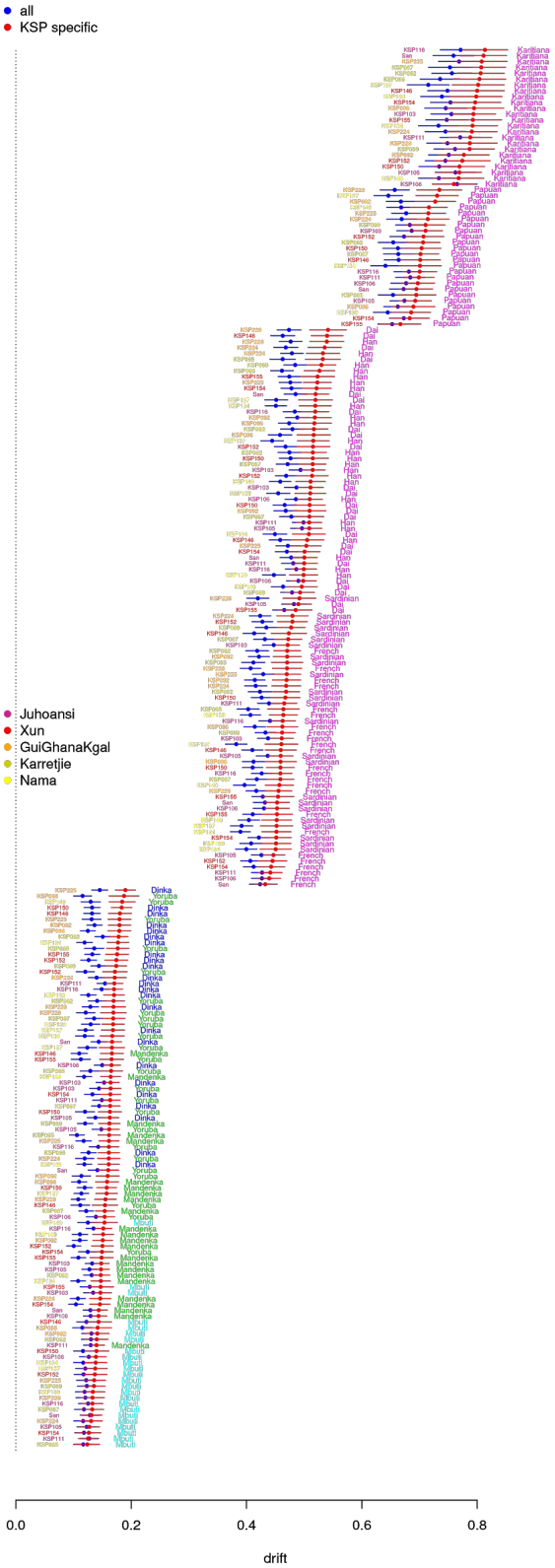


Figure S8.3: Estimated non-Khoe-San branch specific drifts based on the TT method using all data (blue) compared to when restricting the analysis to KSP specific genomic regions (red).

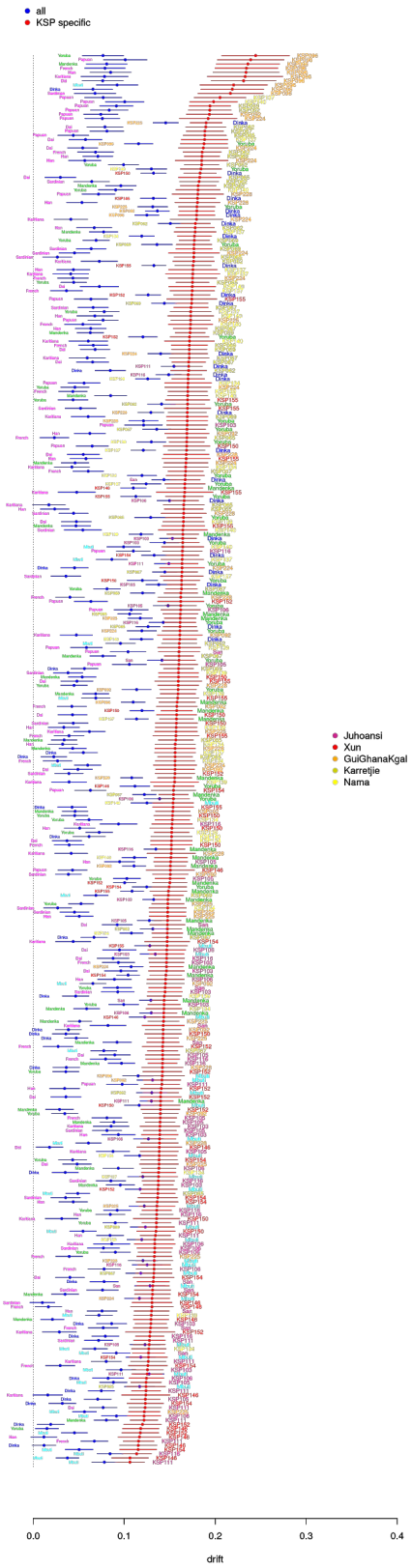


Figure S8.4: Estimated Khoe-San specific branch drifts and other African branch drifts based on the TT method using all data (blue) compared to when restricting the analysis to KSP specific genomic regions (red).

9. MSMC on simulated bottlenecks

9.1 Material and Methods

MSMC (Schiffels and Durbin 2014) can be run on multiple genome sequences. When it is run for two haplotypes, it is similar to PSMC (Li and Durbin 2011). It is then called PSMC' because it differs slightly in the method and underlying model. It is usually run for up to eight haplotypes (four diploid genomes) but can also be used for larger number of haplotypes. Depending on the number of haplotypes, different timeframes can be investigated as the mean first coalescence between two individuals occurs more recently with a larger number of haplotypes. Thus, using eight haplotypes will be more precise for recent times than two haplotypes. Here we investigated what number of haplotypes is optimal for analyzing bottlenecks of varying intensity, duration and starting time.

We simulated haplotypic data under different demographic model using the software MaCS (Chen et al. 2009) and the genetic architecture model from (Beichman et al. 2017) with a few adjustments. We then ran MSMC and plotted the results in R (Team 2013).

9.1.1. Genetic architecture

Each simulated diploid genome consists of 160 chromosomes of length 30 Mbp. This corresponds to a total of 4.8 Gbp which is in the same size order as the diploid human genome. Each 30 Mbp chromosome consists of 300 linked 100-kb recombinating blocks. We used the per-block recombination rates provided by (Beichman et al. 2017) at

https://github.com/LohmuellerLab/Compare_Demographic_Models/tree/master/Simulate_80x30Mb_chromosomes/maCS_RecRatesFiles_Ratios_Feb2016 (based on (Kong et al. 2010; Phung et al. 2016)) as well as the same average recombination rate of $9.257856*10^{-9}$ (Morgan by base pair). We used a mutation rate of $1.25*10^{-8}$ (per base pair per generation) and a generation time of 30 years.

9.1.2. Demographic model

We used a bottleneck model of instantaneous decrease of population size followed by instantaneous recovery as pictured in Figure S9.1. The effective population size before and after the bottleneck, *N*, was fixed to 17,000 individuals. This is a robust estimate for the ancestral population size of humans (Figure S7.3). We chose not to incorporate a recent exponential population growth which is observed in most (but not all) human populations today as we are interested in older events and some of the populations we are interested in, like the Khoe and the San, do not display such a signal of recent population expansion (Mallick et al. 2016).

The three parameters that we vary are:

-$\alpha \in [0.1, 0.2, 0.5, 0.75]$: the intensity of the bottleneck. The effective population size during the bottleneck is αN.

-$T_B \in [15000, 40000, 50000, 250000]$ in years before present: start of the bottleneck.

-$\tau \in [5000, 10000, 20000, 30000]$ in years: duration of the bottleneck. The bottleneck lasts between T_B and T_B-τ years ago.

The combination of parameters that we tested are in Table S9.1. We also tested a null bottleneck model with $\alpha=1$, $T_B=40000$ and $\tau=10000$.

9.1.3. Coalescent simulations with MaCS

We adapted the script provided at

https://github.com/LohmuellerLab/Compare_Demographic_Models/blob/master/Simulate_80x30Mb_chromosomes/SimulateGutenkunst_OutOfAfrica_inMaCS.sh for our own demographic model, mutation rate and recombination rate. We generated a sample of five diploid genomes i.e. ten haplotypes. The MaCS output is piped to "msformatter" (Chen et al. 2009) which

transforms it into Hudons's *ms* format (Hudson 2002); that is then piped into "ms2multihetsep.py" (*https://github.com/stschiff/msmc-tools/blob/master/ms2multihetsep.py*) which generated the MSMC input file for ten haplotypes.

A typical command looked like :

$./macs ${nsam} ${seqLength} -t ${THETA} -r ${RHO} -R ${recRatesFile}/${i}.300blocks.RATIOS.MACS.txt -eN ${TEND} ${ALPHA} -eN ${TB} 1 > ${i}.macsFormat.OutputFile.txt ;

$./msformatter < ${i}.macsFormat.OutputFile.txt | python3 ms2multihetsep.py $i $seqLength > ${i}.msmcFormat.OutputFile.txt ;

with the following values:

nsam=10

seqLength=30000000

THETA=0.00085=4*N*mu with N=17000 and mu=$1.25*10^{-8}$

RHO=0.0006295342=4*N*rho with N=17000 and rho= $9.257856*10^{-9}$

TB=T_B and TEND=T_B-τ scaled in units of 4N generations

ALPHA= α

We then created the input files for two, four or eight haplotypes using custom scripts that 1) selects the relevant genotypes and 2) selects only the variable positions in the subsample. We took all possible combinations of samples, resulting in five "single diploid genomes" inputs, ten "pairs of diploid genomes" inputs, five "four diploid genomes" inputs, and one "five diploid genomes" input.

9.1.4. MSMC runs

We ran a total of 21 MSMC runs for each combination of α, T_B and τ. For the parameter "-r" we used the mutation and recombination rates used in our simulations. Instead of the common value of 0.88 for human data (for example (Schlebusch et al. 2017), we thus used the ratio 0.74. A typical command was:

$msmc --fixedRecombination -r 0.74 -t 3 -o output_msmc/output.ff input/chr*.input.txt;

9.1.5. Plotting

We converted the inverse of the coalescence rate to effective population sizes. We plotted the population size over time in R (Team 2013). We plotted both each of the 21 runs and the average for each number of haplotypes category (i.e. two, four, eight and ten).

9.2 Results

9.2.1. General

The simulation framework is adapted: the ancestral population size is close to the real one. However, we can see from the MSMC run on simulations without a bottleneck (Figure S9.2) that for each number of haplotypes, some time intervals cannot be used. For example for four haplotypes the curve is most accurate between ~200,000 and ~5,000 years ago, and for eight haplotypes between ~10,000 and ~1,000 years ago. For older times, N_e is underestimated.

The results are repeatable: for one combination of parameters two independent MaCS simulations were done. They gave very similar MSMC results.

Sharp changes in population size are seen as progressive (as described in (Schiffels and Durbin 2014)); the decrease in the MSMC curve is longer and less deep than the true trajectory N_e. However, the changes are more abrupt when using only two haplotypes.

We also see that, in general, even if a certain number of haplotypes is more accurate, all number of haplotypes are somehow affected by the bottleneck.

Decrease in estimated N_e due to a bottleneck is sometimes followed by a "bump" (population size larger than true population size). But even without a bottleneck, the curve is not flat.

9.2.2. Recent bottleneck – Example: 15,000 to 5,000 years ago with α=0.1, Figure S9.3

Forward in time, the curves for four, eight or ten haplotypes show a decrease in estimated N_e at the bottleneck followed by recovery. The lowest estimated N_e is reached closer to present that at the actual bottleneck and is more drastic (smaller) than the true value. The estimated Ne-trajectory is closer to the true trajectory using four haplotypes than using eight and ten haplotypes. With two haplotypes, there is a drop from the ancestral population size a bit before the true bottleneck and no recovery (until 1,000 years ago at least). The same observations hold for a longer bottleneck (duration of 10,000 years) except that that the true bottleneck N_e is reached (figure not shown).

9.2.3. Intermediate bottleneck (starting 50,000 years ago), Figures S9.4 to S9.6

The "bottleneck shape" (decrease followed by recovery) is clearer for two or four haplotypes than for eight or ten haplotypes. For eight and ten haplotypes, the curves start around 100,000 years ago and thus do not extend very far into the period of time pre-dating the bottleneck (that ends at 80,000 years ago). This is likely the reason why the ancestral size is not recovered in these trajectories. The more haplotypes, the more recent the implied bottleneck. The curves for two haplotypes are particularly accurate when the bottleneck is long and intense (see for example Figure S9.4, α=0.1 and bottleneck during 30,000 years) but fails to recover more recent population sizes. For less intense bottlenecks (see for example Figure S9.5, α=0.2 and duration 5,000 years) the "bottleneck shape" is not as clear for eight and ten haplotypes as the curves start from a small population size. For even weaker bottlenecks (see for example Figure S9.6, α=0.5 and duration 5,000 years), the "bottleneck shape" is seen for four haplotypes only.

9.2.4. Ancient bottleneck – Example: 250,000 to 230,000 years ago with α=0.1, Figure S9.7

The "bottleneck shape" is seen for two haplotypes only. The duration of the bottleneck and N_e during the bottleneck are overestimated. Eight and ten haplotypes are not relevant as the investigated time period is too ancient. The curves with four haplotypes show an increase in population size from a low point of about 8,000 individuals (similar to lowest point estimated with two haplotypes) to a quite accurate recent population size.

9.3 Discussion

Based on these simulations results, we see that for a basic bottleneck model, the most accurate results are obtained for:

Recent bottleneck (15,000 years ago): four, eight or ten haplotypes.

Intermediate bottleneck (50,000 years ago): two or four haplotypes.

Ancient bottleneck (250,000): two haplotypes.

A bottleneck signal is visible using MSMC with two, four, eight or ten haplotypes except in very extreme cases (very recent and weak bottleneck, or very ancient) and there is no case where only one of the MSMC curve shows a signal.

In some specific cases (Figure S9.6) the "bottleneck shape" is seen mostly with four haplotypes.

We were not able to recover the observed N_e trajectories using our empirical data (Figure S7.12) where a drop in N_e starting 100,000 years ago and until ~10,000 years ago is observed for four and eight haplotypes but not for two. This is perhaps not surprising considering the very simple

demographic scenarios simulated here. For example, we might need to include population (sub)structure or/and admixture from a population which experienced a bottleneck. However, our analysis shows that it is entirely possible that a (true) bottleneck is detected by MSMC only for a specific number of haplotypes and not for any other number of haplotypes.

MSMC on simulated bottlenecks Table and Figures

Table S9.1: Combinations of parameters tested.

τ \ T_B	**15000**	**40000**	**50000**	**250000**
2500			α=0.05, 0.1, 0.5	
5000	α=0.1	α=0.2	α=0.05, 0.1, 0.2, 0.5, 0.75	α=0.1
10000	α=0.1	α=0.2	α=0.05, 0.1, 0.2, 0.5, 0.75	α=0.1
20000	-	α=0.2	α=0.05, 0.1, 0.2, 0.5, 0.75	α=0.1
30000	-	-	α=0.05, 0.1, 0.2, 0.5, 0.75	-

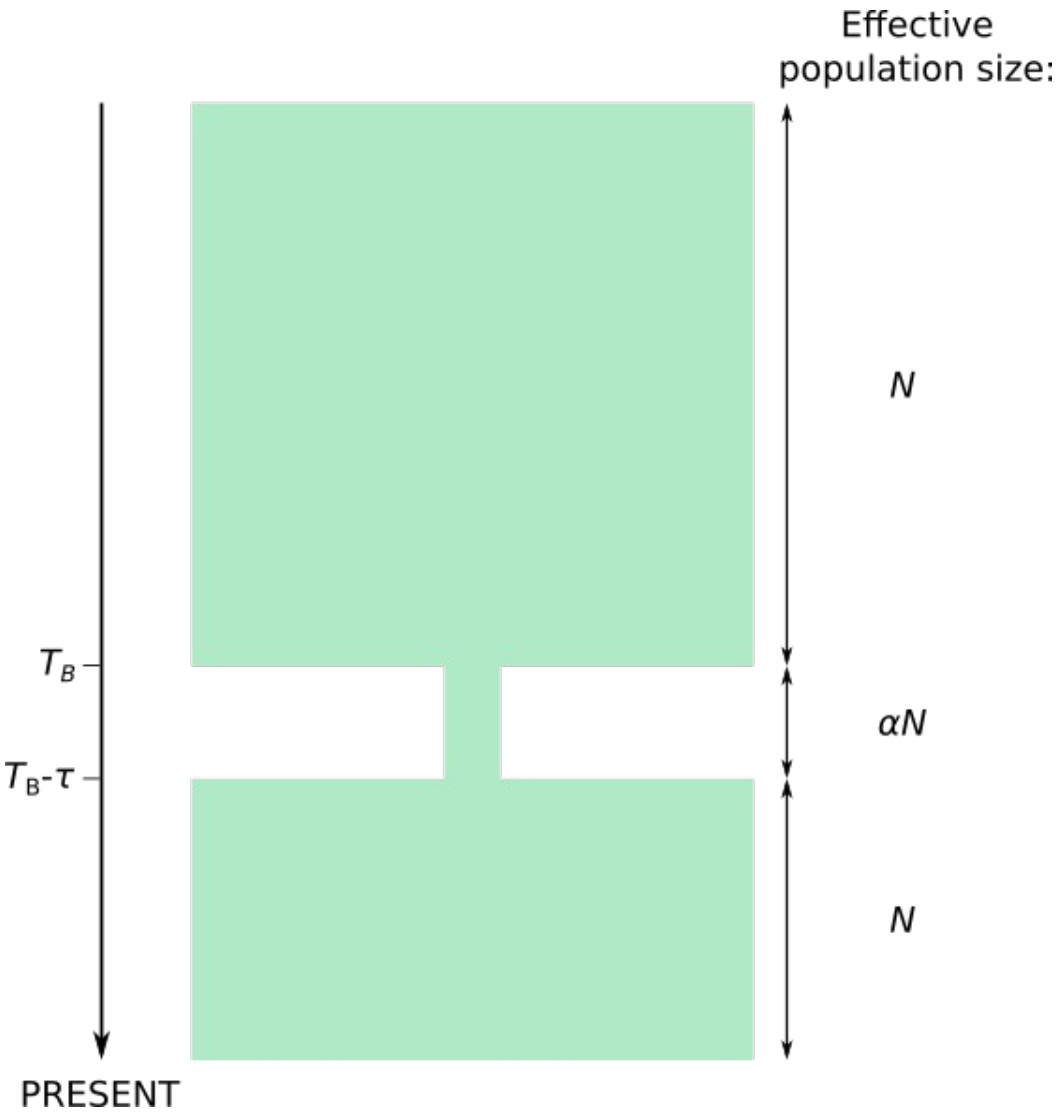


Figure S9.1: Demographic model used in the simulations.

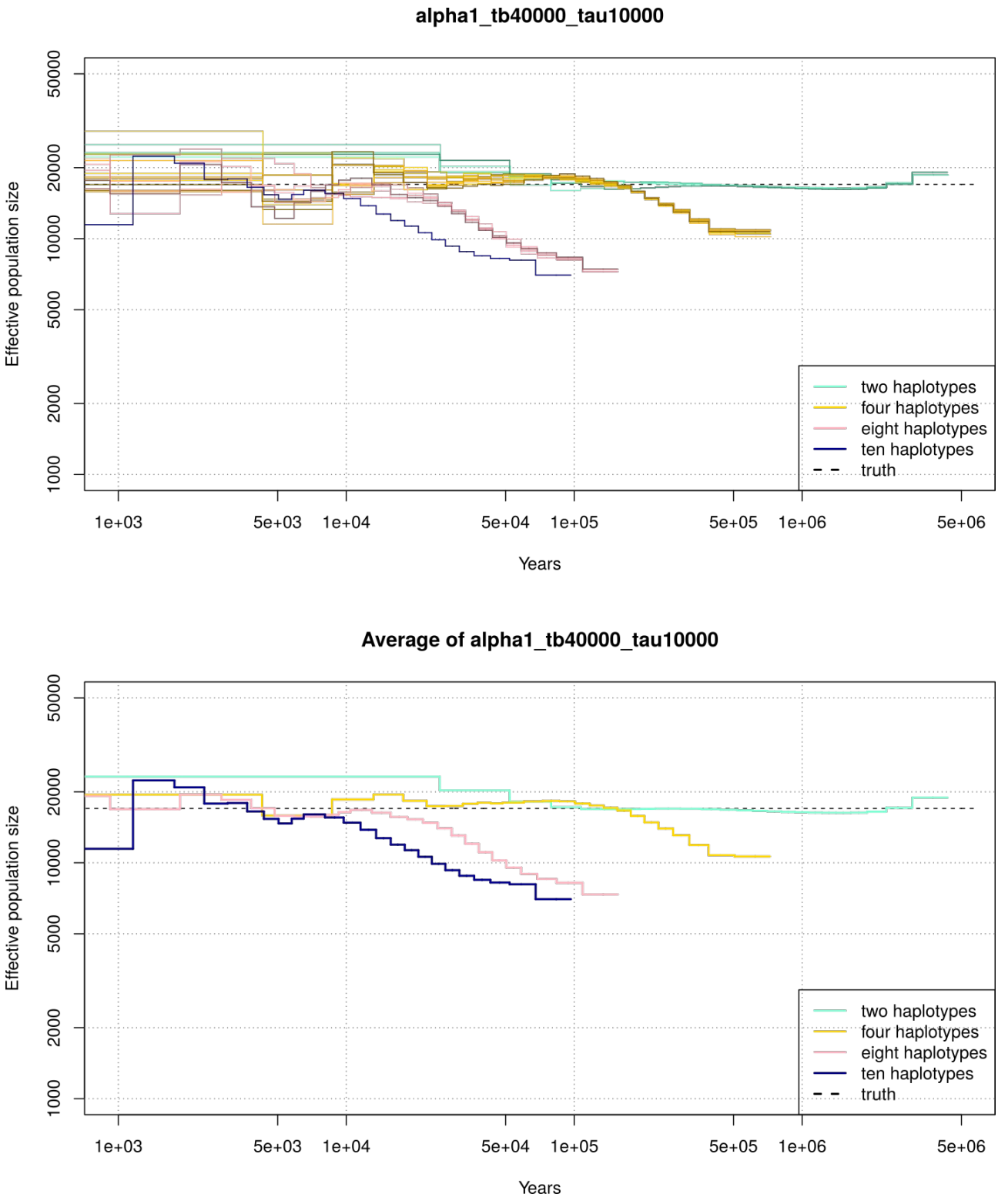


Figure S9.2: MSMC curves (solid lines) and true *N*e trajectory (dashed line). Logscale: x and y axis. Demographic model: No bottleneck. α=1.

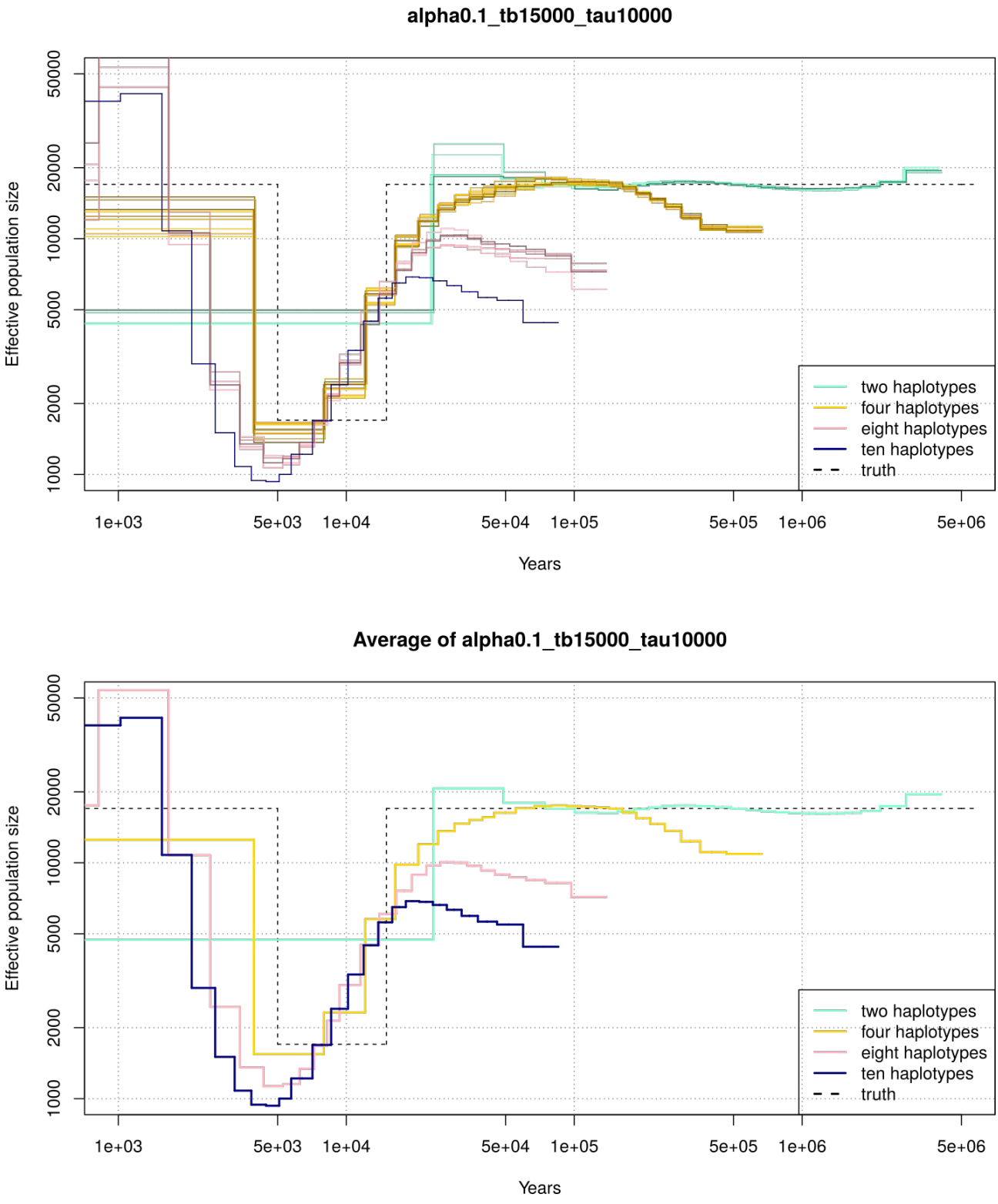


Figure S9.3: MSMC curves (solid lines) and true *N*e trajectory (dashed line). Logscale: x and y axis. Demographic model: Recent bottleneck. α=0.1, T_B=15,000, τ=10,000.

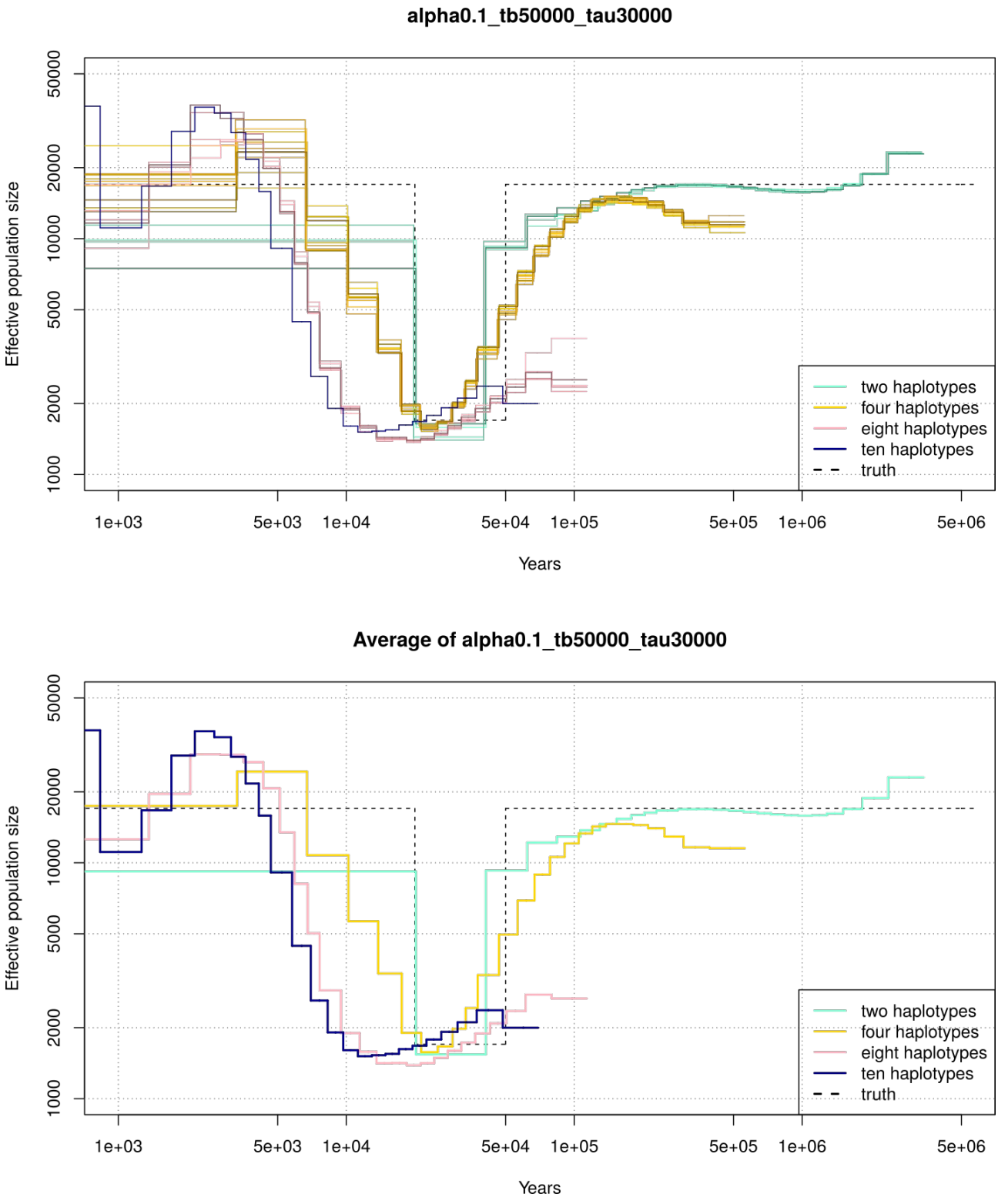


Figure S9.4: MSMC curves (solid lines) and true *N*e trajectory (dashed line). Logscale: x and y axis. Demographic model: Intermediate strong bottleneck. α=0.1, T_B=50,000, τ=30,000.

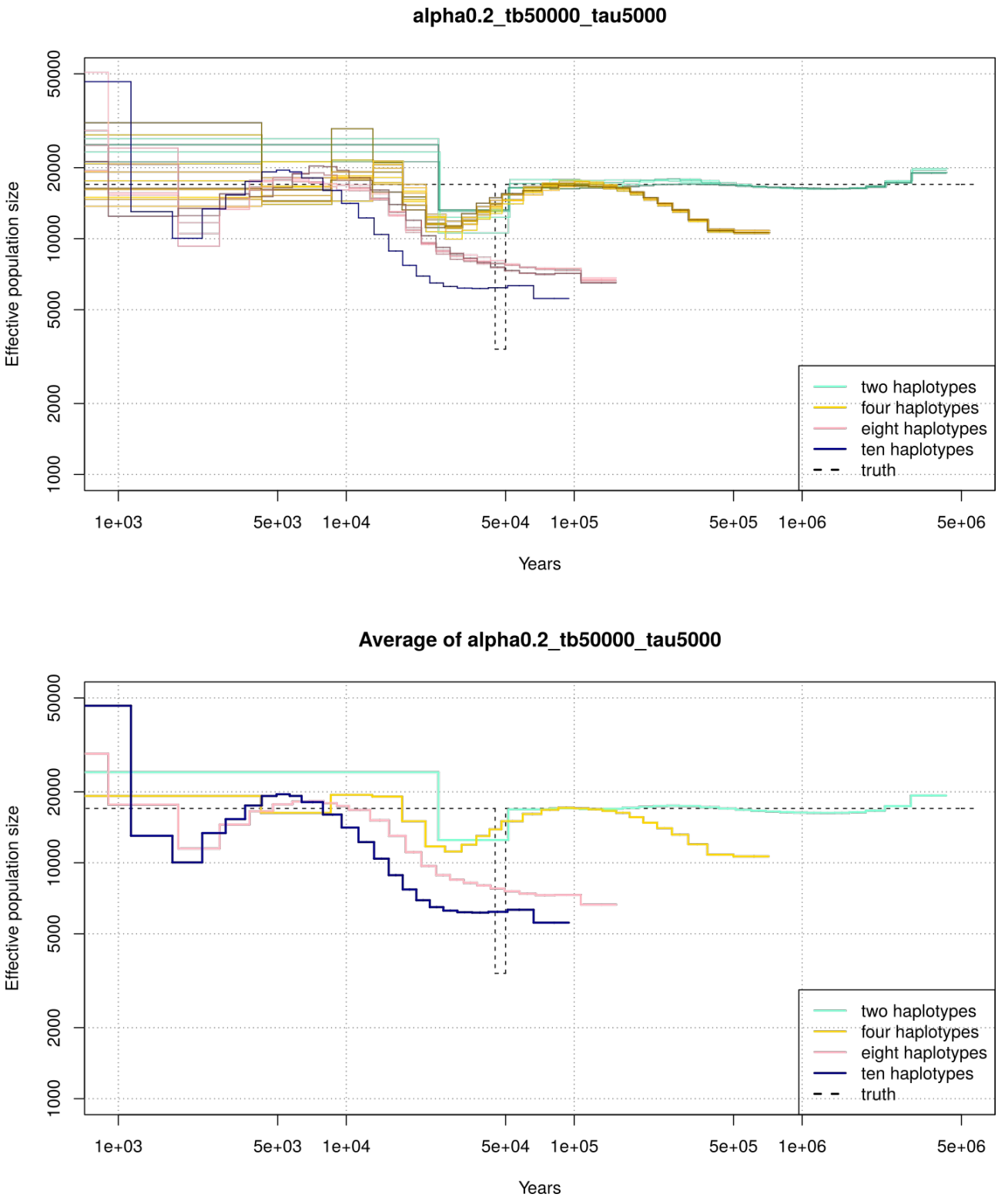


Figure S9.5: MSMC curves (solid lines) and true *N*e trajectory (dashed line). Logscale: x and y axis. Demographic model: Intermediate medium bottleneck. α=0.2, T_B=50,000, τ=5,000.

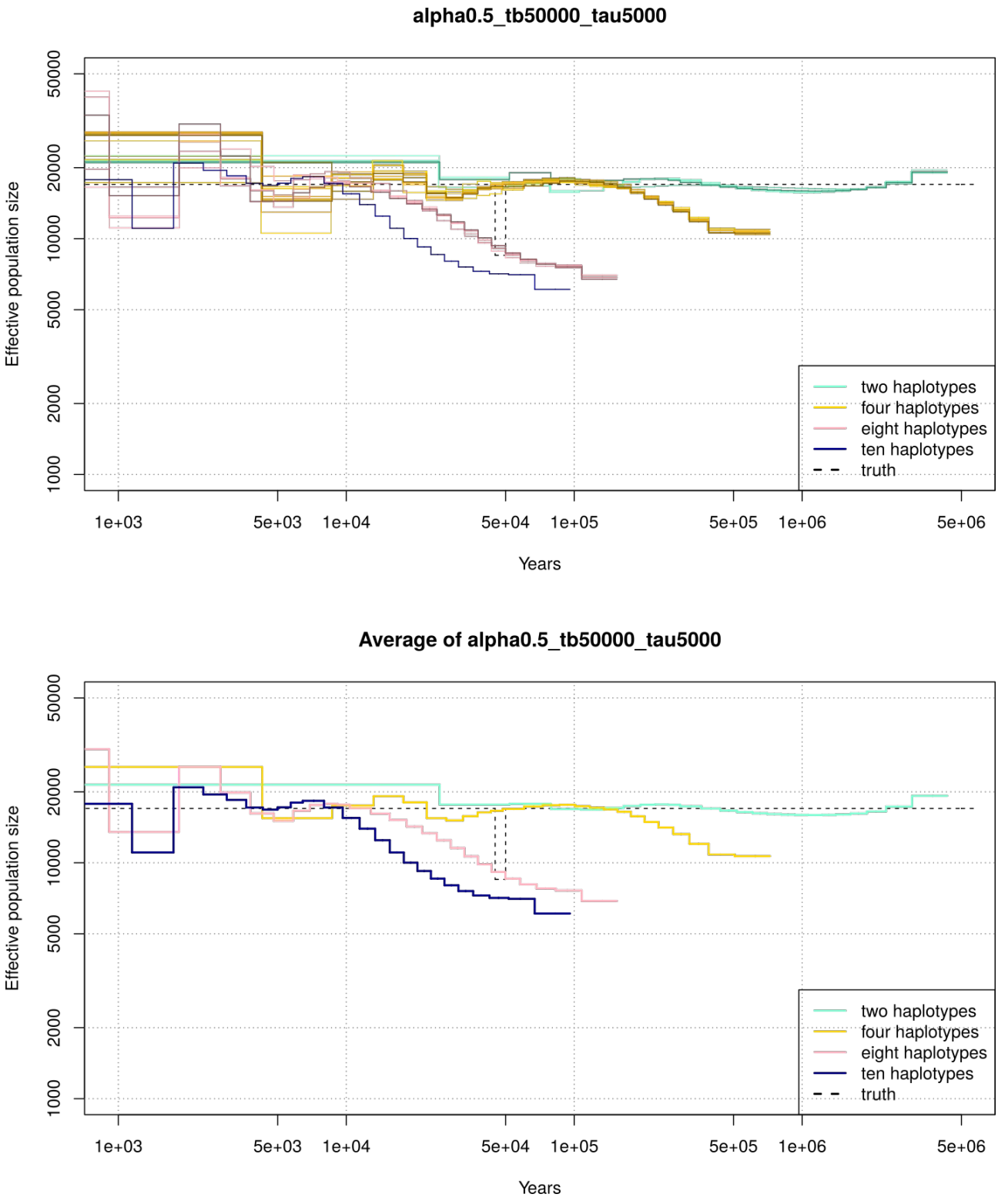


Figure S9.6: MSMC curves (solid lines) and true *N*e trajectory (dashed line). Logscale: x and y axis. Demographic model: Intermediate weak bottleneck. α=0.5, T_B=50,000, τ=5,000.

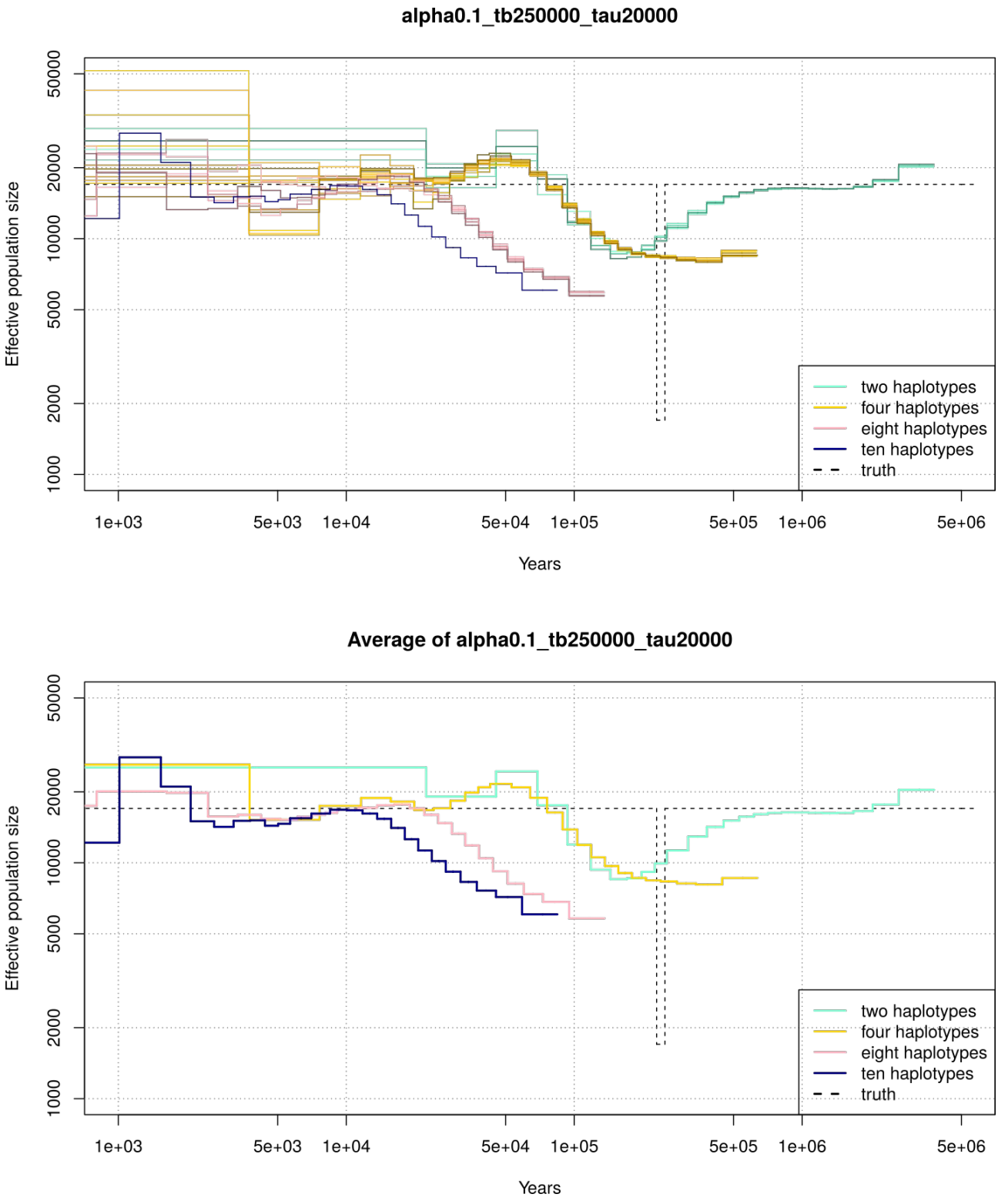


Figure S9.7: MSMC curves (solid lines) and true *Ne* trajectory (dashed line). Logscale: x and y axis. Demographic model: Old strong bottleneck. α=0.1, T_B=250,000, τ=20,000.

10. Inference of archaic admixture with *S**

We used the *S** (Plagnol and Wall 2006) statistic to investigate the proposed archaic admixture event in Africans (Plagnol and Wall 2006; Wall et al. 2009; Lachance et al. 2012). Based on *S**, previous research have suggested that especially the western rainforest hunter-gatherer groups display genetic signatures of archaic admixture (Plagnol and Wall 2006). Since lineage sorting is typically common with putative archaic sources (such as the Neandertals and Denisovans), a reference population that is assumed not to share the archaic admixture event is an efficient way to decrease shared ancestry due to lineage sorting: SNPs for which the derived allele is observed in the reference population are not considered. To search for archaic admixture in African populations it is natural to use the non-African populations as a reference population. This excludes the Neandertals and Denisovans as a source of archaic admixture in Africa which seems like a valid assumption since the technique was clearly successful in identifying the Neandertal and Denisovan introgression in non-Africans using African populations as a reference and also because there is a striking homogeneity of D (when used to test for either Neandertal or Denisovan introgression (Figures S6.8-S6.13) or for a relative difference of Neandertal vs Denisovan introgression (Figures S6.14-15)) across African populations.

We find, however, that *S** is sensitive to both sample size of the reference population and the relative "phylogenetic" position of the investigated populations. The distribution of *S** using only the 11 HGDP individuals with five African individuals and six non-African populations is the most right-skewed for the San individual while the other four African individuals show lower values of *S** in general (Figure S10.1 top left panel). In contrast, when we include the additional 25 KSP-individuals in the analysis, the Mbuti show the most right-skewed distribution while all the KSP-individuals (including the HGDP-San individual) have lower *S** values (Figure S10.1 top right panel).

This analysis also reveals the importance of a good reference population because the Denisovan admixture event in the Papuan individual is considerably more visible when all 30 African individuals are included (Figure S10.1 top right panel) compared to when only the five African HGDP individuals are included (Figure S10.1 top left panel).

Performing the analysis exclusively on the HGDP-San and the HGDP-Mbuti individuals using the six non-African HGDP individuals as reference gives an *S** distribution that is considerably higher for the San than the Mbuti individual (Figure S10.1 middle left panel). We interpret this result as being due to the closer relationship between the Mbuti individual and the non-African individuals compared to the relationship between the San individual and the non-African individuals. In fact, the *S** distribution is essentially identical between the San and the Mbuti individual when no reference population is used (Figure S10.1 middle right panel). However, the highest *S** values are in this case found in Mbuti (Figure S10.1 bottom left) which may be indicative of archaic admixture (see also (Lachance et al. 2012) for western rainforest hunter-gatherer groups) but at this stage we consider this relatively weak evidence, and we see no strong evidence for assuming that there has been archaic introgression in any of the Khoe-San populations. Future aDNA studies on African populations might make inferences of archaic admixture in African populations a more testable hypothesis - as it did for non-Africans - but with current genetic data and tools we could not find any conclusive evidence of admixture with archaic forms of humans in African populations.

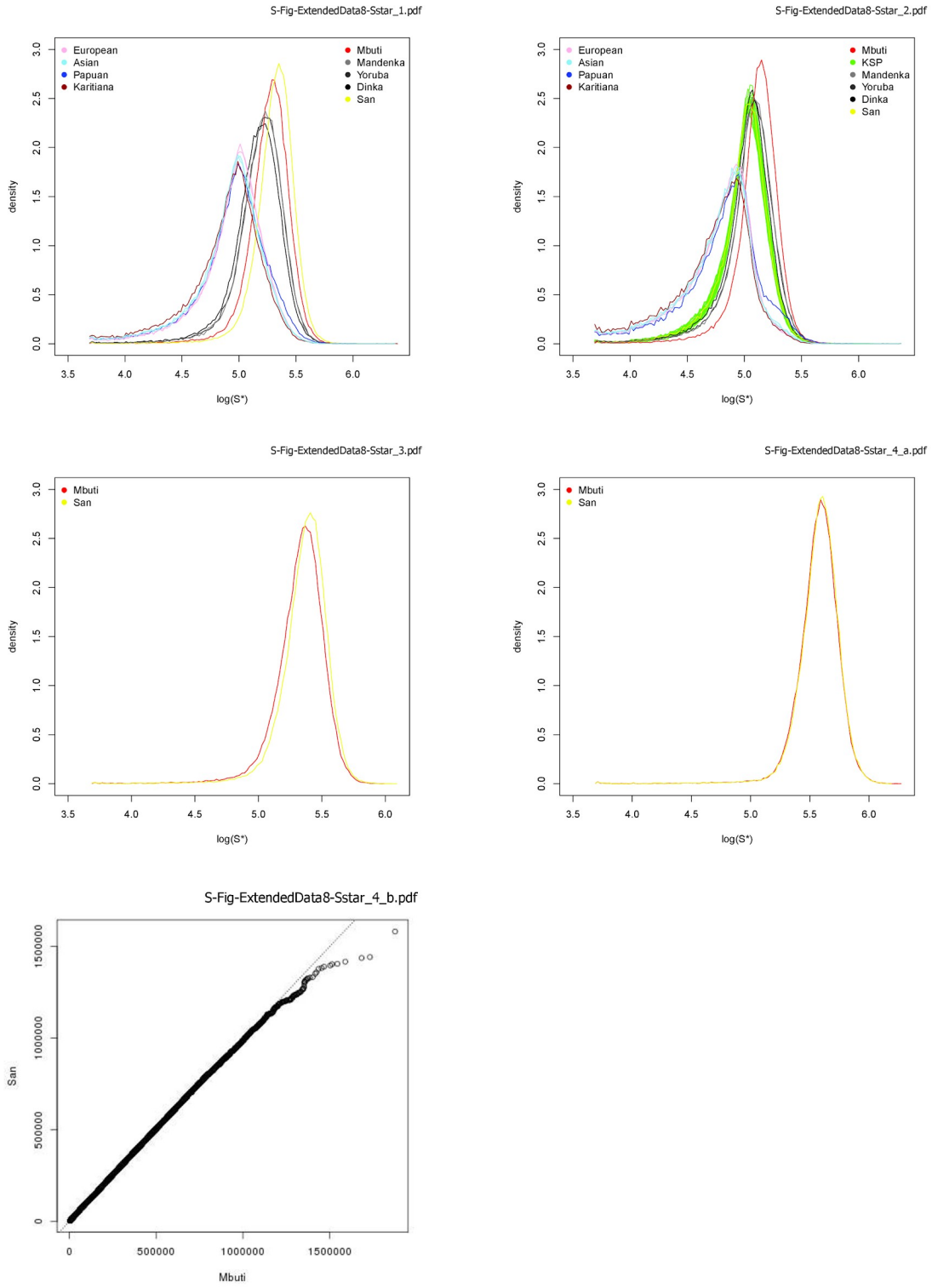


Figure S10.1: Inference of archaic admixture with *S**. Top left panel: inferences with HGDP individuals only (six non-African reference individuals and five African individuals tested). Top right panel: inference with KSP+HGDP individuals (total of 30 African individuals). Middle left panel: contrast HGDP San and HGDP Mbuti with the six non-African HGDP samples. Middle right panel: contrast HGDP San and HGDP Mbuti without reference population. Bottom left panel: San against Mbuti.

11. Selection scans

We looked at candidate regions in iHS and XP-EHH scans in various grouping levels of African groups. For iHS, ten candidate regions were considered per population according to the highest ratio of variants with |iHS|>2 in 100 kb windows. For XP-EHH, we considered ten candidate regions according to the most extreme value of XP-EHH in each population.

11.1. iHS computation

iHS was computed using selscan program developed by Z. Szpiech (Szpiech and Hernandez 2014) with default parameters. We then normalised iHS values using the program "norm" associated with selscan and considered the absolute value. We averaged over windows of length 100,001 bp with a step length of 1 kb between windows. We then only considered windows with (strictly) more than 50 SNPs. We calculated iHS in various groups of human populations at different regional scales.

We first investigated possible signals of selection in Khoe-San using individuals available from this study (five !Xun, five Ju|'hoansi, five Nama, five Karretjie, five |Gui and ||Gana), resulting in a sample size of 25 individuals with 7,451,085 SNPs variants.

Within the Khoe-San dataset, northern Khoe-San and southern Khoe-San were analyzed separately. We computed iHS in northern San on a sample consisting of five !Xun and five Ju|'hoansi individuals for 9,463,168 SNPs. iHS calculations in southern San, were based on a sample of five Nama and five Karretjie individuals. For the southern Khoe-San, iHS could be calculated for 9,527,765 SNP variants.

We furthermore searched for signals of selection among all non Khoe-San Africans present in the "Global dataset". This sample includes individuals from Dinka, Mbuti, Yoruba, Mandenka, Hadza, Yoruba from Ibadan (YRI), Luhya in Webuye (LWK), Maasai in Kinyawa (MKK), Baka rainforest hunter-gatherers, Bakola rainforest hunter-gatherers, Bedzan rainforest hunter-gatherers and Sandawe. In total, 33 non Khoe-San individuals, referred to as "other Africans", were grouped into a single sample and iHS was calculated for 5,725,019 SNP variants.

11.2. XP-EHH computation

XP-EHH was also computed using selscan (Szpiech and Hernandez 2014) with default parameters. We normalised the XP-EHH values using the program "norm" associated with selscan with default parameters. We then average over windows of length 100,001 bp with a step length of 1 kb between windows. The windows considered afterwards were required to contain strictly more than 50 SNPs.

First we computed XP-EHH comparing northern Khoe-San versus southern Khoe-San, each group had a sample size of ten individuals and XP-EHH could be calculated for a total of 13,519,185 variants. Within Africa we then look at signals specific to Khoe-San and non Khoe-San groups by calculating XP-EHH in the 26 Khoe-San individuals versus the 33 non Khoe-San individuals, for 21,297,406 variants.

11.3. Selection of candidate regions

We selected the top 1% of windows with the highest |iHS|>2 ratio. We then merged overlapping windows within this 1% into larger regions where each region was associated with a value corresponding to the highest |iHS|>2 ratio among its constituent windows. We considered as candidates the ten regions with the highest associated |iHS|>2 ratio. A similar merging of XP-EHH of windows based on the top 1% of windows with the most extreme (either negative or postive) XP-EHH value was performed.

Figure S11.1 shows the Manhattan plots of ratio |iHS|>2 in northern and southern Khoe-San as well as for all Khoe-San together. It also shows the position of their respective ten candidate regions. Figure S11.2 shows the Manhattan plots of standardized XP-EHH in northern and southern Khoe-San as well as for all Khoe-San together.

11.4. Enrichment analysis using GOWINDA

We used Gowinda (Kofler and Schlötterer 2012) to perform unbiased gene-set enrichment of Gene Ontology terms (GO-terms) among outliers of |iHS| and XP-EHH as they are SNP-based. The GO-term list was downloaded from the funcAssociate database (Berriz et al. 2009), and we used the Human genome annotation from ENSEMBL build 37, release 75 (Flicek et al. 2013).

For iHS, we selected the variants with the top 1% highest |iHS| values as candidate SNPs. For each XP-EHH scan, we consider both top 1% highest positive values as being candidates for selection in one population and the top 1% lowest negative values as being candidate for selection in the other population (non-Africans were not considered for selection and only used as a reference population).

We ran GOWINDA with default parameters, using gene mode and considering, as belonging to genes, any variants within +-500 bp of coding regions. GO-terms are considered as enriched if they show an FDR lower than 10 %.

11.5. Adaptation in northern San

Northern San are hunter-gatherers, who supposedly kept the proto Khoe-San lifestyle. They are also less admixed with non-Khoe San populations than southern San and the study of the adaptation patterns in their genome could tell us about adaptation that has been going on in Khoe-San for longer times.

Among the ten candidate regions that we identified as potential targets of selection in northern Khoe-San (Table S11.1), we discuss below some that are particularly interesting regarding the genes they overlap.

On chromosome 10, the region between 47.5 Mb and 47.6 Mb, shows a strong signal of selection. It contains a pseudogene, the “anthrax toxin receptor-like pseudogene 1” (*ANTXRLP1*). The highest |iHS| value falls in *ANTXRLP1*, at position 47,607,481. About 40 kb from this region is also the gene “anthrax toxin receptor-like” (*ANTXRL*). These two sequences show similarity to the anthrax toxin receptors to which the toxin of the anthrax bacteria *Bacillus anthracis* binds and enters the host cell. Anthrax is endemic to Namibia and affects mostly wild animals but also humans (Turner et al. 2013). *Bacillus anthracis* can infect humans when in close contact with wild animals or livestock, after the ingestion of undercooked contaminated meat, as well as it can be transmitted from infected animals to humans by blood feeding insects (Turner 1980; Turell and Knudson 1987). Also, *Bacillus anthracis* spores can be found in the soil and on the grass around a carcass of an infected animal up to two years after the death (and the soil around a carcass gets more fertile so probably an area favored for gathering) (Turner et al. 2013).

Another interesting region of potential adaptation in northern San is found on chromosome 11 between 58.3 Mb and 58.8 Mb. The highest |iHS| value in this region is found approximately 30 kb upstream of the *GLYAT* gene, on position 58,445,133. The *GLYAT* gene codes for a mitochondrial protein that is involved in detoxifying xenobiotics but also endogenous organic acids. The highest |iHS| value being approximately 30 kb from this gene could be located in a regulating region for the expression of this gene.

Within the class I MHC region on chromosome 6, there is an |iHS| signal around 29.8 Mb and 30 Mb. The highest per-SNP |iHS|-value of this region is intronic to *HLA-G*. HLA-G is involved in the presentation of foreign antigens to the immune system and plays a role in maternal tolerance of the fetus by mediating its protection from the deleterious effects of natural killer cells, cytotoxic T-lymphocytes, macrophages and mononuclear cells. *HLA-G* is expressed in fetal derived placental cells. It has been recently shown (although for other genes) (Hilton et al. 2015), that Khoe-San probably have a better tolerance during pregnancy than other human groups, protecting against abortion and preeclampsia.

When looking for signals of selection specific to northern San when compared to southern San using XP-EHH (Table S11.2), two of the highest XP-EHH are found in two candidate regions in the MHC complex on chromosome 6. The first is located between 31.2 Mb and 31.6 Mb and has its highest XP-EHH variant intronic to *HLA-B*, the second region is located between 32.1 Mb and 32.5 Mb with the highest XP-EHH variant in *C6orf10*. The MHC region is known to be highly selected and divergent between populations (Traherne 2008), and it is often among the candidate regions for selection.

On chromosome 8 between 121.7 Mb and 121.9 Mb, there is an XP-EHH peak in northern San compared to southern San on the gene *SNTB1*, coding for "syntrophin, beta 1". Syntrophin beta 1 is associated with dystrophin and dystrophin-related proteins. Interestingly syntrophin beta 1 interacts with the viral HTLV-1 TAX protein. The human HTLV-1 virus is endemic almost all around the globe but not necessarily affecting all ethnic groups similarly (Gessain and Cassar 2012). Its epidemiology is rather poorly documented and the only study from southern Africa showed that its prevalence is relatively small (0.2% of the studied sample) in Africans from the Pretoria area in South Africa (Goubau et al. 1993) , but the ethnicity is not documented in this study.

On chromosome 1 from 100.2 Mb to 100.4 Mb, we observed an interesting peak specific to northern San compared to southern San. The highest XP-EHH hit of this region is intergenic, upstream of two genes, *FRRS1* and *AGL*. *FRRS1* codes for ferric-chelate reductase 1, which reduces ferric to ferrous iron before its transport from the endosome to the cytoplasm. It has been associated with dietary absorption of iron (McKie et al. 2001). *FRRS1* is expressed in the stomach, duodenum, small intestine and liver, strengthening its potential role in the dietary intake of iron. The other gene in this region, *AGL* coding for amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase, or the glycogen debrancher enzyme, is involved in glycogen degradation. Mutations in the *AGL* gene are associated with glycogen storage disease III, a metabolic disorder due to an accumulation of abnormal glycogen in the liver, muscles and, in some cases, the heart. It is clinically characterized by hepatomegaly, hypoglycemia, short stature, and variable myopathy. This disease is cured by high protein diet.

The signals of putative selection signals we identify in the northern Khoe-San is consistent with selection acting on diet, pathogens and --indirectly to pathogens-- better pregnancy tolerance.

11.6. Adaptation in southern San

In the southern San (Nama & Karretjie People), the Nama are characterized by a pastoralist diet that was probably adopted from a hunter-gatherer lifestyle around 2,000 years ago. The diet for southern African Khoe pastoralists (such as the Nama) has been heavily dairy based with only incidental consumption of meat, mainly from hunting (Lombard and Parsons 2015). They present complex patterns of admixture from other African and non-African groups.

Among the ten candidate regions selected according the ratio of |iHS|>2 (Table S11.3), the highest |iHS| hit in southern San is located on chromosome 12 from 20.9 Mb to 21.1 Mb. The highest variant (21,012,422) falls into an intron of *SLCO1B3* also overlapping a sequence of the non-functional gene *SLCO1B7*. *SLCO1B3* codes for "solute carrier organic anion transporter family, member 1B3", it is liver-specific and mediates the sodium-independent uptake of endogenous and xenobiotic compounds. It also plays a critical role in bile acid and bilirubin transport. Bilirubin is a product of degradation of hemoglobin, it is taken-up by the liver where it is conjugated with glucuronic acid, making it soluble in water. Much of the bilirubin goes into the bile. Bile is secreted by the liver to the small intestine and helps the digestion of fats. There is thus a potential link between this gene and a change to a pastoralist diet with increased fats from dairy and meat products. This gene has also been associated with height (Gudbjartsson et al. 2008).

The second highest |iHS| value is located on chromosome 6 at 26,409,106 in a region from 26.3 Mb to 26.7 Mb. This region is located close to the MHC region and is also among the candidate regions for XP-EHH in Khoe-San versus other Africans. It contains many genes, mostly tRNA genes but also several genes of the butyrophilin family (BTN). The highest |iHS| signal is intronic to *BTN3A1*.

The butyrophilins are immunoglobins involved in the immune system and also involved in the fat droplet secretion in milk (Abeler-Dörner et al. 2012). Butyrophilins has been suggested to have a protective function for the infant as they are transmitted via the milk (Peterson et al. 2001).

Another candidate region was found on chromosome 11 between 120.8 Mb and 121.0 Mb. The highest |iHS| value of this region (120,892,212) falls 2.5 kb upstream of the *TBCEL* gene. *TBCEL* codes for "tubulin folding cofactor E-like protein", which acts as a regulator of tubulin stability. This protein is abundantly expressed in testis, but is also present in several tissues at a much lower level. It has been suggested to be involved in male fertility (Nuwal et al. 2012) which is a major selective target in men.

When looking at signals of selection that are specific to southern Khoe-San compared to northern Khoe-San (Table S11.4), the most extreme signal is located on chromosome 6 between 54.2 Mb and 54.4 Mb. The highest XP-EHH variant is intergenic about 55 kb upstream to *TINAG*, coding for the "tubulointerstitial nephritis antigen". This gene encodes a glycoprotein located in the kidney, but also expressed in the small intestine and cornea. Autoantibodies against this protein are found in sera of patients with various nephritis and nephropathy, all causing renal damages. This gene seems to be regulated in a precise spatial and temporal pattern throughout nephrogenesis (formation and development of the kidney in the foetus) (Kanwar et al. 1999).

The second strongest signal of selection specific to southern Khoe-San compared to northern Khoe-San is found on chromosome 2 between 25.7 and 26.0 Mb, with the highest value at position 25,873,841, intronic to the gene *DTNB*. This gene encodes dystrobrevin beta, a component of the dystrophin-associated protein complex (DPC). DPC disruption is associated with various forms of muscular dystrophy. *DTNB* is expressed in the brain, the kidney and the pancreas. Interestingly, dystrobrevin beta is thought to interact with syntrophins and a syntrophin coding gene, *STNB1*, is among the candidate regions in northern Khoe-San.

The widest region among the candidate regions for selection in southern Khoe-San compared to northern Khoe-San, is located on chromosome 9 from 88.4 Mb to 88.7 Mb. This region also corresponds to a candidate region for southern Khoe-San based on iHS (Table S11.3). The highest |iHS| SNP variant (6.67014) is intronic to *NAA35*, whereas the highest XP-EHH SNP variant is 70 kb upstream from *NAA35*, in *AK124523*. *NAA35* codes for an auxiliary component of the N-terminal acetyltransferase C (NatC) complex which catalyzes acetylation of N-terminal methionine residues. *NAA35* is involved in regulation of apoptosis and proliferation of smooth muscle cells. A strong |iHS| signal is this region overlaps part of the gene *GOLM1* which is upregulated in cases of viral infection. Although the function of this gene is unknown, the protein encoded by this gene is a type II Golgi transmembrane protein. This gene is widely expressed, predominantly in epithelia. Interestingly, this gene is expressed at low level in normal liver, and its expression is significantly higher in hepatitis B and C infected liver, not in liver disease due to non-viral causes. Hepatitis B has an important prevalence in all of Sub-Saharan Africa whereas Hepatitis C is more dependent on countries and has a rather low prevalence in southern African countries (except Zimbabwe).

A last region indicating stronger selection in southern Khoe-San than in northern Khoe-San, is located on chromosome 4 between 70.8 and 71.0 Mb with the highest variant at position 70,905,193**.** The highest XP-EHH variant falls between the gene *HTN3* (5kb downstream) and *HTN1* (10 kb upstream), both coding for the precursor of all the histatin proteins of the histatin family. Histatins function as antimicrobial and antifungal peptides and are important components of the innate immune system. Histatins are found in saliva and function in wound healing. Another salivary protein coding gene in the region is *STATH* (statherin) that belongs to the same family as histatins. Statherin stabilizes saliva supersaturated with calcium salts by inhibiting the precipitation of calcium phosphate salts. It also modulates hydroxyapatite crystal formation on the tooth surface. Two functional (*CSN2* and *CSN1S1*) and two pseudogenes (*CSN1S2AP* and *CSN1S2BP*) of casein genes are also found in this region. Beta casein (encoded by *CSN2*) is the principal protein in human milk and the primary source of essential amino acids for a suckling infant. The alpha s1 casein

(encoded by *CSN1S1*) plays an important role in the capacity of milk to transport calcium phosphate. The two pseudogenes are pseudogenes of the casein alpha s2.

In summary and similar to northern Khoe-San, immunity is implicated as a selective target in the southern Khoe-San. However, some selection on immunity may in addition be linked to a pastoralist diet. The selection on biliburin production and milk components is potentially an adaptation to diet.

11.7. Adaptation in Khoe-San

We also searched for signals of selection within in all Khoe-San groups considered together (Table S11.5).

The region with the highest ratio of |iHS|>2 in Khoe-San is located on chromosome 1, between 202.3 Mb and 202.5 Mb with the highest per-SNP |iHS| value intronic to the gene *PPP1R12B*. This gene is coding for the subunit 12B of the protein phosphatase 1. This protein has for function to regulate the myosin phosphatase activity and augments Ca(2+) sensitivity of the contractile apparatus. The *PPP1R12B* gene is expressed in skeletal muscle, fetal and adult heart, brain, placenta, kidney, spleen, thymus, pancreas and lung. This selection signal is also strong in southern Khoe-San. As well as dystrobrevin and syntrophin showing signals of selection in northern Khoe-San and southern Khoe-San respectively, this gene is also important in the functioning of skeletal muscle.

Approximately three megabases before this region, between 219.2 Mb and 219.4 Mb, is a region with up to 61% of variants with |iHS|>2, which is also one of the strongest signals for |iHS| in northern Khoe-San. The highest |iHS| variant of this region is around 15 kb upstream of the gene *LYPLAL1* (lysophospholipase-like 1). *LYPLAL1* has been positively associated with adiponectin hormone level (Dastani et al. 2012), adiposity (Lindgren et al. 2009), and Waist-Hip Ratio (Heid et al. 2010). *LYPLAL1* has been associated with sexual dimorphism, especially is has a strong female-only association with fat distribution (Lindgren et al. 2009; Heid et al. 2010). *LYPLAL1* has also been associated with fasting insulin levels (Scott et al. 2012).

On chromosome 4, between 17.7 Mb and 18.1 Mb, we located a candidate selected region in Khoe-San with up to 62% of |iHS|>2. The highest |iHS| variant is found at position 18,029,162, approximately 7 kb upstream of the *LCORL* gene. This region is furthermore found to have relatively high XP-EHH values in Khoe-San compared to other Africans, with the highest XP-EHH variant 30 kb upstream of *LCORL*. *LCORL* (ligand dependent nuclear receptor corepressor-like) encodes a transcription factor that appears to function in spermatogenesis. Polymorphisms in this gene are also associated with skeletal frame size and adult height (Gudbjartsson et al. 2008; Soranzo et al. 2009; N'Diaye et al. 2011; Carty et al. 2012). *LCORL* overlaps with the gene *NCAPG*, which has also associated with height (Gudbjartsson et al. 2008) and is coding for the condensin subunit G, which is highly expressed in testis (among the genes in the ten candidate regions, *TNFRSF10D* is also expressed highly in the testis).

Using XP-EHH to search for potentially selected regions specific to Khoe-San compared to other Africans (Table S11.6), two more regions are implicated.

On chromosome 5, we find a selection signal specific to Khoe-San compared to other Africans between 77.7 Mb and 77.9 Mb with the highest XP-EHH variant at position 77,767,577, intronic to *SCAMP1*. *SCAMP1* codes for the secretory carrier membrane protein 1 that functions as carrier to the cell surface in post-golgi recycling pathways. This gene is widely expressed, with highest expression in brain, and has been positively associated with hippocampal atrophy (Potkin et al. 2009). Hippocampal atrophy, is the atrophy of the hippocampus, a part of the brain believed to be involved in memory. Another gene found in this region is *LHFPL2*, lipoma HMGIC fusion partner-like 2, a gene with potential involvement in benign lipoma tumors and deafness.

Another interesting region is on chromosome 15 from 26.7 Mb to 26.9 Mb, with the highest XP-EHH variant located at 26,778,527. This variant is 20 kb downstream of the gene *GABRB3*, coding

for the “gamma-aminobutyric acid (GABA) A receptor, beta 3”, a member of the ligand-gated ionic channel family. This protein is one of the subunits of a chloride channel functioning as receptor for GABA. The gene may be associated with the pathogenesis of several disorders including Angelman syndrome (DeLorey et al. 1998), Prader-Willi syndrome (Cassidy 1997), autism (Kim et al. 2006) and cognitive performance in general (Need et al. 2009). Interestingly, both Angelman and Prader-Willi syndromes, although being pathogenic states, are characterised --among other traits-- by a hypopigmented skin and a short stature. GABRB3 also functions as histamine receptor and mediates cellular responses to histamine.

11.8. Adaptation in other Africans

By investigating the |iHS| and XP-EHH scans in non Khoe-San Africans, we might get insights of selection happening after the divergence between Khoe-San and non Khoe-San Africans (Table S11.7 and 11.8). Non Khoe-San Africans live in very diverse environments with very diverse lifestyles. Figure S11.3 shows the Manhattan plot of ratio |iHS|>2 in non-Khoe-San Africans with the position of the ten candidate regions.

The second highest ratio of |iHS|>2 in non Khoe-San Africans is found on chromosome 1 between 27.3 Mb and 27.6 Mb with the highest |iHS| variant at position 27,455,505 intronic to *SLC9A1*. *SLC9A1* is part of the solute carrier family and codes for NHE1, cation proton antiporter 1, involved in pH regulation to eliminate acids generated by active metabolism or to counter adverse environmental conditions. This gene is specifically expressed in kidney and intestine.

On chromosome 6 between 56.9 Mb and 57.1 Mb there is another potentially selected region in non Khoe-San Africans, with the highest |iHS| variant located at 57,041,124, intronic to the gene *BAG2*. Another gene close to the highest |iHS| variant is *RAB23*, that encodes a small Rab GTPase of the Ras superfamily. Rab proteins are involved in the regulation of diverse cellular functions associated with intracellular membrane trafficking, including autophagy and immune response to bacterial infection.

In non Khoe-San Africans, the highest XP-EHH value is located on chromosome 10 at position 89,256,573 in a candidate region between 88.9 Mb and 89.4 Mb. The highest XP-EHH variant is located approximately 15 kb upstream from the gene *MINPP1* coding for multiple inositol polyphosphate phosphatase; which removes 3-phosphate from inositol phosphate substrates. It is the only enzyme known to hydrolyze inositol pentakisphosphate and inositol hexakisphosphate (IP6 or phytic acid). Phytic acid is a saturated cyclic acid and the principal storage form of phosphorus in many plant tissues, especially bran and seeds. It can be found in cereals and grains. Phytic acid is not digested by humans but it tends to decrease mineral (such as calcium, iron, and zinc) or vitamin (such a niacin) absorption from the food, as it chelates minerals and vitamins. The deficiency of niacin causes pellagra and lack of minerals can cause various pathologies. An efficient phytic acid degradation can be therefore advantageous in a vegetable-rich diet, especially agriculturalist diet. The selected variant is upstream to the gene *MINPP1*, suggesting that it could modify its level of expression and act on its enzymatic activity in certain tissues. A recent study showed that IP6 is absent from human serum compared to other species suggesting that *MINPP1* is secreted and degrades IP6 in order to avoid its mineral-chelating effect (Wilson et al. 2015). The group of non Khoe-San Africans in this study, contain many groups with farming lifestyles, while none of the Khoe-San groups have adopted crop farming as a lifeway.

11.9. Enrichment signals

Enrichment analysis showed that many of the observed selection signals were associated with DNA modification and transcription related GO-terms. It is indeed the most numerous genes in the genome (Mi and Thomas 2009). The enrichment of these genes is hard to interpret in terms of biological effect as the genes in these GO-terms can have broad functions.

Among other GO-terms with a more specific biological function and FDR<10% (Table S11.9), we find some interesting GO-terms. In northern Khoe-San for example, the GO-term “response to

arsenic-containing substance" is enriched among |iHS| outliers. There are indications of arsenic in drinking water resources due to mining activity in southern Africa, but the concentration of water sources is poorly documented. GO-terms related to immunity are found in southern Khoe-San and non Khoe-San. Response to xenobiotics, such as those found in the candidate regions are also enriched in XP-EHH outliers for Khoe-San compared to non Khoe-San. We also observe, enrichment in XP-EHH outliers in olfactory receptors for both Khoe-San and non Khoe-San testifying the fast adaptive evolution of the olfactory receptor genes reported several times in several populations (Gilad et al. 2003; Nielsen et al. 2007; Duforet-Frebourg et al. 2016).

Selection scans Tables and Figures

Table S11.1: Detail of the ten candidate regions identified as potential targets of selection in northern Khoe-San, using the ratio of variant with |iHS|>2 in 100 kb windows. The position of the highest |iHS| value over the region was used to identify the closest gene.

Chr	Start	End	Max |iHS|>2 ratio	Max iHS value	Position of maximum | iHS| SNP	Genes	Gene closest to max(iHS)
6	29880000	30075000	0.795	5.44	29958676	HLA-J ZNRD1 HLA-H HCG4B HLA-A UNQ6501 HLA-G PPP1R11 ZNRD1-AS1 RNF39 TRIM31 HCG9	intronic to HLA-G, HLA-H & HLA-J
1	219223000	219460000	0.768	5.054	219326970	LYPLAL1 LYPLAL1-AS1	in LYPLAL1-AS1, 20kb upstream of LYPLAL1
6	109787000	110019000	0.657	5.367	109869484	MICAL1 FIG4 ZBTB24 AKD1	intronic to AK9
11	31456000	31679000	0.648	5.169	31582231	IMMP1L ELP4	intronic to ELP4
8	33783000	34040000	0.638	4.877	33969051	no genes	no genes
11	58275000	58811000	0.625	7.052	58445133	CNTF OR5B21 LPXN GLYATL2 GLYATL1 AB231721 GLYAT ZFP91	30kb downstream of GLYAT
10	59856000	60116000	0.622	3.738	60020982	IPMK CISD1 UBE2D1	intronic to IPMK
10	47520000	47631000	0.621	7.936	47607481	ANTXRLP1	in ANTXRLP1
10	32509000	32710000	0.607	4.974	32636859	EPC1	intronic to EPC1
7	141872000	142069000	0.592	7.275	141952582	PRSS58, MOXD2P TRYP2 MGAM2	intronic to PRSS58

Table S11.2: Ten candidate regions for XP-EHH in northern San compared to southern San, based on the most extreme XP-EHH variants. The position of the most extreme XP-EHH value is used to consider the closest genes.

Chr	Start	End	Position of max	Max XP-EHH	Genes	Closest gene to max(XP-EHH)
1	100161000	100365000	100265337	26.87	AGL FRRS1	between FRRS1 & AGL
21	15636000	15837000	15737111	26.88	HSPA13 ABCC13	10kb downstream of HSPA13
8	119806000	120013000	119913562	27.12	TNFRSF11B	20kb downstram of TNFRSF11B
8	121672000	121882000	121771519	27.43	SNTB1	intronic to SNTB1
14	56505000	56706000	56606829	28.00	PELI2	intronic to PELI2
4	17597000	17801000	17698457	29.08	MED28 LAP3 FAM184B	intronic to FAM184B
5	49552000	49755000	49647411	30.15	EMB	centromeric
6	31185000	31576000	31287379	31.89	HCG26 DDX39B LST1 HLA-C HLA-B NCR3 LTA MCCD1 LTB HCP5 TNF ATP6V1G2 NFKBIL1 PMSP MICB MICA	intronic to HLA-B
11	42594000	42950000	42708516	40.94	No genes	no genes
6	32139000	32501000	32268980	47.95	HLA-DRA BTNL2 HCG23 NOTCH4 PPT2 HLA-DRB1 AGER HLA-DRB5 AGPAT1 C6orf10 RNF5 PBX2 GPSM3	in C6orf10

Table S11.3: Detail of the ten candidate regions identified as potential targets of selection in southern Khoe-San, using the ratio of variant with |iHS|>2 in 100 kb windows. The position of the highest |iHS| value over the region was used to identify the closest gene.

Chr	Start	End	Max |iHS|>2 ratio	Max iHS value	Position of maximum | iHS| SNP	Genes	Gene closest to max(iHS)
6	27864000	28415000	0.918	7.611	28084732	ZNF192 ZNF193 TRNA_Ser TRNA_Gly TOB2P1 PGBD1 NKAPL ZNF165 ZSCAN23 ZNF187 ZNF323 ZSCAN12P1 TRNA_Met ZSCAN12 ZKSCAN3 ZKSCAN4 OR2B2 BC043177 OR2B6 ZSCAN16	10kb upstream of ZSCAN16
6	26268000	26699000	0.707	8.611	26409106	TRNA_Ile HMGN4 BTN3A2 BTN3A3 TRNA_Ser BTN3A1 HIST1H4H BTN2A3P BTN2A2 BTN2A1 TRNA_Tyr TRNA_Gln BTN1A1 TRNA_Thr TRNA_Arg TRNA_Lys ZNF322 HIST1H3G HIST1H2BI TRNA_Met TRNA_Leu TRNA_Trp TRNA_Ala HCG11 ABT1 TRNA_Val TRNA_Pro	intronic to BTN3A1
9	88420000	88728000	0.662	6.67	88595276	GOLM1 NAA35	intronic to NAA35
11	120806000	121014000	0.621	5.294	120892212	TECTA TBCEL GRIK4	2.5kb upstream of TBCEL
4	4151000	4265000	0.615	5.67	4246433	OTOP1 TMEM128	intronic to TMEM128, 20kb upstream of OTOP1
1	202303000	202510000	0.596	4.869	202326027	U6 UBE2T PPP1R12B	intronic to PPP1R12B
12	20901000	21126000	0.583	9.499	21012422	SLCO1B3 SLCO1B7 SLCO1C1	intronic to SLCO1B3
5	27906000	28151000	0.576	5.822	28062724	no genes	no genes
2	112946000	113152000	0.572	5.744	113047540	ZC3H8 RGPD8 RGPD5 RGPD6 ZC3H6	intronic to ZC3H6

6	32190000	32442000	0.541	6.394	32325399	HLA-DRA HCG23 NOTCH4 C6orf10 BTNL2	in C6orf10

Table S11.4: Ten candidate regions for XP-EHH in southern San compared to northern San, based on the most extreme XP-EHH variants. The position of the most extreme XP-EHH value is used to consider the closest genes.

Chr	Start	End	Position of max	Max XP-EHH	Genes	Closest gene to max(XP-EHH)
6	54200000	54422000	54311642	-36.1259	TINAG	intergenic
2	25772000	25980000	25873841	-31.7116	ASXL2 DTNB	intronic to DTNB
2	86949000	87139000	87049389	-26.4509	CD8A CD8B RMND5A	intronic to CD8B
7	143680000	143885000	143785994	-26.2369	ARHGEF35 OR2A12 CTAGE4 OR2A14 OR2A25 OR2A2 OR2A5 OR6B1	7kb upstream from OR2A12
8	131685000	131884000	131784694	-25.5001	ADCY8	10kb downstream from ADCY8
1	72290000	72489000	72389068	-25.0988	NEGR1	intronic to NEGR1
6	30605000	30871000	30771037	-25.06	MIR4640 DDR1 C6orf136 Nbla00487 NRM MDC1 AK098012 TUBB IER3 DHX16 FLOT1 PPP1R18 ATAT1	15kb downstream from LINC00243
22	18845000	19041000	18940214	-24.9694	DGCR9 PRODH DGCR11 Y_RNA DGCR2 DGCR5 DGCR6 DGCR10	15kb upstream of both PRODH and DGCR5
4	70806000	71005000	70905193	-23.882	CSN1S2BP HTN3 HTN1 CSN2 CSN1S1 CSN1S2AP STATH	5kb downstram from HTN3
9	88387000	88746000	88489520	-23.4423	GOLM1 NAA35	70kb from NAA35

Table S11.5: Detail of the ten candidate regions identified as potential targets of selection in all Khoe-San together, using the ratio of variant with |iHS|>2 in 100 kb windows. The position of the highest |iHS| value over the region was used to identify the closest gene.

Chr	Start	End	Max |iHS|>2 ratio	Max iHS value	Position of maximum |iHS| SNP	Genes	Gene closest to max(iHS)
1	202303000	202512000	0.799	4.676	202408502	U6 UBE2T PPP1R12B	intronic to PPP1R12B
1	3791000	3908000	0.725	6.113	3814290	C1orf174 DFFB LOC100133612	intronic to C1orf174
2	112944000	113159000	0.707	6.145	113047540	ZC3H8 ZC3H6 RGPD6 RGPD5 RGPD8 FBLN7	intronic to ZC3H6
8	22895000	23119000	0.694	6.492	23013880	TNFRSF10C TNFRSF10B TNFRSF10A MGC31957 TNFRSF10D CHMP7 LOC286059 LOC389641	intronic to TNFRSF10D
5	49582000	49837000	0.673	6.226	49652526	EMB	in centromere
14	97942000	98175000	0.667	5.119	98086835	LOC100129345	15kb downstream of LOC100129345
19	37056000	37276000	0.663	5.68	37171380	ZNF382 ZNF529 AX747375 BC024306 ZNF850 ZNF567 ZNF461 BC039524	7kb upstream of ZNF567
4	17735000	18121000	0.625	4.835	18029162	NCAPG DCAF16 FAM184B LCORL	7kb upstream of LCORL
1	219238000	219440000	0.613	5.357	219335220	LYPLAL1 LOC643723	15kb upstream of LYPLAL1, in LOC643723
1	87002000	87220000	0.597	5.614	87160987	CLCA3P SH3GLB1 CLCA4	about 10kb upstream of SH3GLB1

Table S11.6: Detail of the ten candidate regions for XP-EHH in Khoe-San compared to non Khoe-San Africans, based on the most extreme XP-EHH variants. The position of the most extreme XP-EHH value is used to consider the closest gene.

Chr	Start	End	Position of max	Max XP-EHH	genes	Closest gene to max(XP-EHH)
6	32195000	32468000	32399158	16.4469	C6orf10 BTNL2 HLA-DRA HCG23	10kb upstream of HLA-DRA
12	26054000	26269000	26168989	16.4572	RASSF8	intronic to RASSF8
6	81215000	81421000	81315596	16.9937	no_genes	no genes
1	238462000	238663000	238562534	17.3681	no gene	intergenic
15	26675000	26909000	26778527	17.4711	GABRB3	10kb downstream of GABRB3
3	93659000	93852000	93750941	18.0233	ARL13B NSUN3 PROS1 DHFRL1 STX19 U7	intronic to ARL13B
13	80357000	80556000	80457220	18.2641	BC036310	in BC036310
6	26309000	26637000	26428702	18.5842	TRNA_Ile HMGN4 BTN3A2 TRNA_Gln TRNA_Ser BTN3A1 BTN2A3P BTN2A2 BTN2A1 TRNA_Tyr BTN3A3 BTN1A1 TRNA_Thr TRNA_Arg TRNA_Lys ZNF322 TRNA_Met TRNA_Leu TRNA_Trp TRNA_Ala HCG11 ABT1 TRNA_Val TRNA_Pro	intronic to BTN2A3P pseudogene
5	77664000	77870000	77767577	18.8125	LHFPL2 SCAMP1	intronic to SCAMP1
6	27863000	28079000	27973364	19.182	TRNA_Met ZNF165 TRNA_Gly ZSCAN12P1 OR2B2 OR2B6	50kb downstream of OR2B6

Table S11.7: Detail of the ten candidate regions identified as potential targets of selection in other Africans (non-Khoe-San), using the ratio of variant with |iHS|>2 in 100 kb windows. The position of the highest |iHS| value over the region was used to identify the closest gene.

Chr	Start	End	Max |iHS|>2 ratio	Max iHS value	Position of maximum | iHS| SNP	Genes	Gene closest to max(iHS)
15	74728000	74971000	0.781	4.159	74824123	UBL7 ARID3B EDC3 CLK3	10kb upstream of ARID3B
1	27282000	27596000	0.78	4.831	27455505	TRNP1 SLC9A1 WDTC1 FAM46B C1orf172	intronic to SLC9A1
16	25597000	25748000	0.731	5.278	25653146	HS3ST4	50kb upstream of HS3ST4
8	53076000	53299000	0.728	4.724	53229227	ST18	intronic to ST18
13	68247000	68470000	0.723	4.126	68378293	no genes	no genes
12	109847000	110052000	0.7	5.794	109920993	UBE3B MVK KCTD10 MMAB MYO1H	intronic to UBE3B
12	82554000	83159000	0.696	5.42	82866149	CCDC59 TMTC2 C12orf26	intronic to METTL25
10	93121000	93312000	0.692	6.778	93217772	HECTD2	intronic to HECTD2
7	87205000	87590000	0.689	5.89	87528502	SLC25A40 DBF4 ABCB1 RUNDC3B ADAM22	intronic to DBF4
6	56904000	57119000	0.686	5.889	57041124	BAG2 ZNF451 RAB23	intronic to BAG2

Table S11.8: Detail of the ten candidate regions for XP-EHH in non Khoe-San Africans compared to Khoe-San, based on the most extreme XP-EHH variants. The position of the most extreme XP-EHH value is used to consider the closest gene.

Chr	Start	End	Position of max	Max XP-EHH	Genes	Closest gene to max(XP-EHH)
10	88883000	89360000	89256573	-40.1035	FAM22D FAM22A MINPP1 FAM35A	7kb upstream of MINPP1
10	51403000	51588000	51488478	-30.5952	PARG NCOA4 TIMM23B MSMB AGAP7 TIMM23	intronic to PARG/TIMM23
6	32379000	32627000	32493077	-28.4896	HLA-DRA HLA-DRB1 HLA-DRB6 HLA-DRB5 HLA-DQA1	intronic to HLA-DRB5
10	17705000	17945000	17777703	-27.9798	TMEM236 STAM MRC1	between STAM and TMEM236
17	20122000	20367000	20238341	-27.1166	CCDC144C SPECC1 LGALS9C LGALS9B	intronic to CCDC144CP
4	34644000	34857000	34754859	-24.5542	no genes	no genes
10	50949000	51198000	51057720	-24.0296	PARG OGDHL	intronic to PARG
17	44690000	44894000	44756501	-23.529	WNT3 NSF	intronic to NSF
19	42420000	42633000	42526054	-22.3036	RABAC1 GRIK5 ARHGEF1 ZNF574 ATP1A3 POU2F2	intronic to GRIK
17	16513000	16799000	16693395	-21.9736	ZNF624 USP32P1 CCDC144A FAM106CP KRT16P2	intronic to USP32P1

Table S11.9: Some GO-terms enriched in |iHS| or XP-EHH outliers in the various African groups. In parenthesis is the total amount of GO-terms presenting an FDR lower than 10%. Note that no GO-terms were enriched for XP-EHH outliers specific to northern San compared to southern San.

	Population	GO-term	FDR	Description
|iHS|	Northern San (33)	GO:0046685	0.053	response to arsenic-containing substance
		GO:0070875	0.068	positive regulation of glycogen metabolic process
		GO:0060740	0.069	prostate gland epithelium morphogenesis
	Southern San (23)	GO:0019884	0.092	antigen processing and presentation of exogenous antigen
	Khoe-San (22)	GO:0002199	0.073	zona pellucida receptor complex
	non KS (292)	GO:0016032	0.004	viral process
		GO:0019012	0.023	virion
		GO:0044403	0.025	symbiosis, encompassing mutualism through parasitism
		GO:0002474	0.033	antigen processing and presentation of peptide antigen via MHC class I
XP-EHH	SS vs. NS (12)	GO:0010043	0.078	response to zinc ion
	KS vs. Non KS (20)	GO:0004984	0.008	olfactory receptor activity
		GO:0001533	0.020	cornified envelope
		GO:0009410	0.036	response to xenobiotic stimulus
	Non KS vs. KS (7)	GO:0004984	0.007	olfactory receptor activity

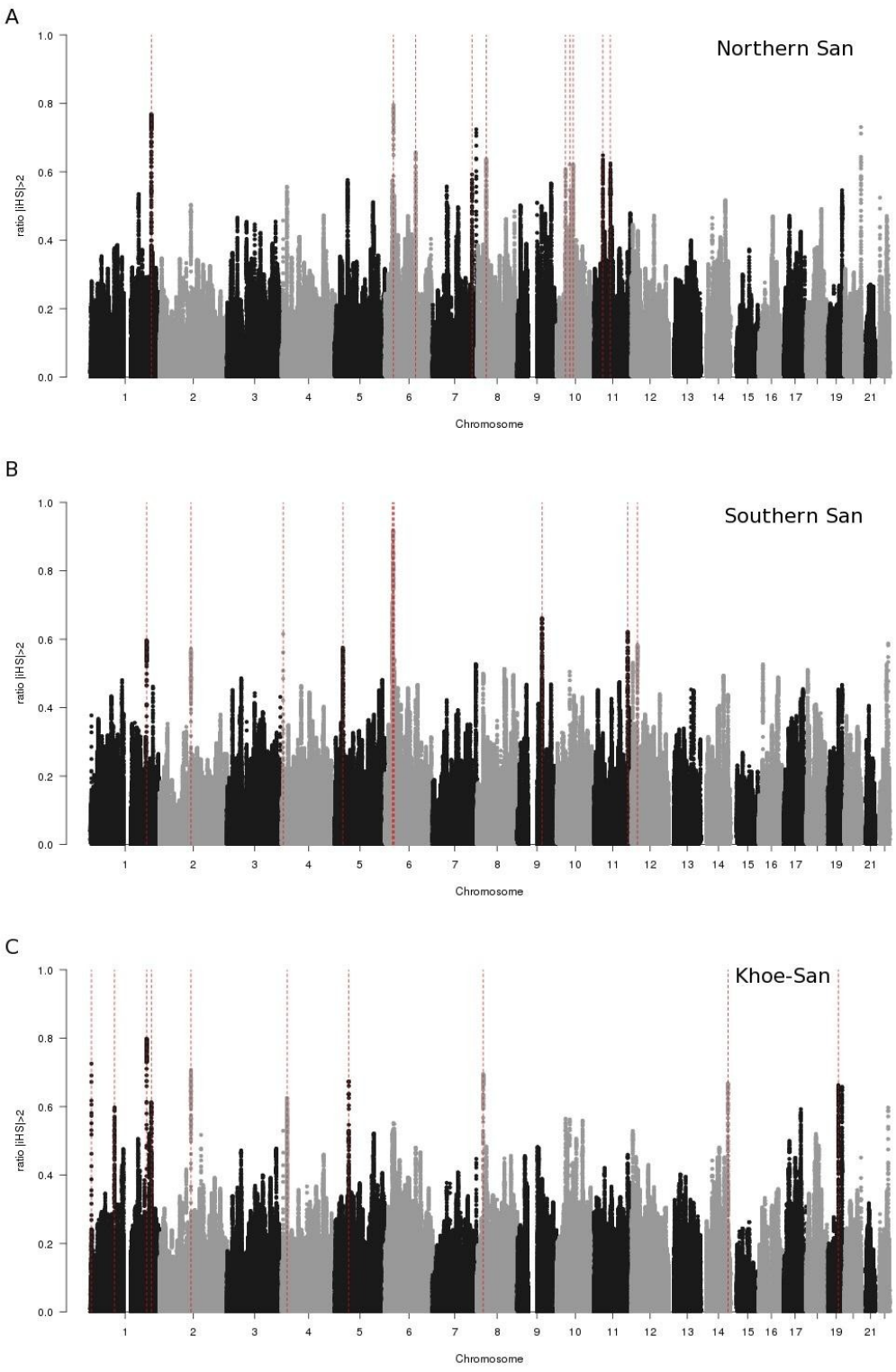


Figure S11.1: Manhattan plot of ratio |iHS|>2 in Khoe-San groups. (A) Northern Khoe-San, (B) Southern Khoe-San, (C) Khoe-San. Red dotted lines mark the position of the position of the highest |iHS| value in the ten candidate regions.

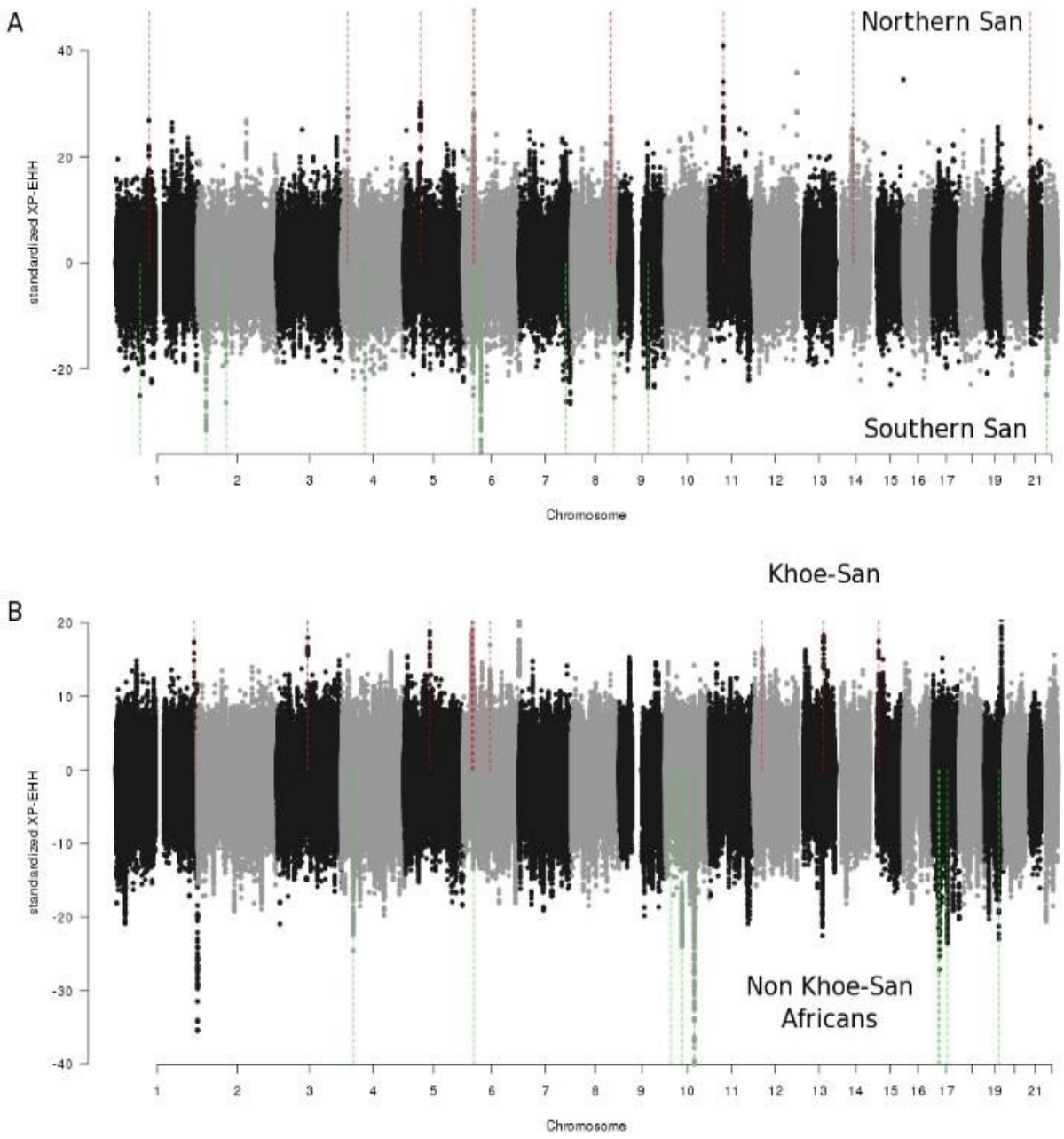


Figure S11.2: Manhattan plot of standardized XP-EHH. (A) In northern Khoe-San vs. southern Khoe-San. (B) In Khoe-San vs. non Khoe-San Africans. Red dot lines mark the position of the ten highest positive values, green dotted lines mark the position of the ten most negative values.

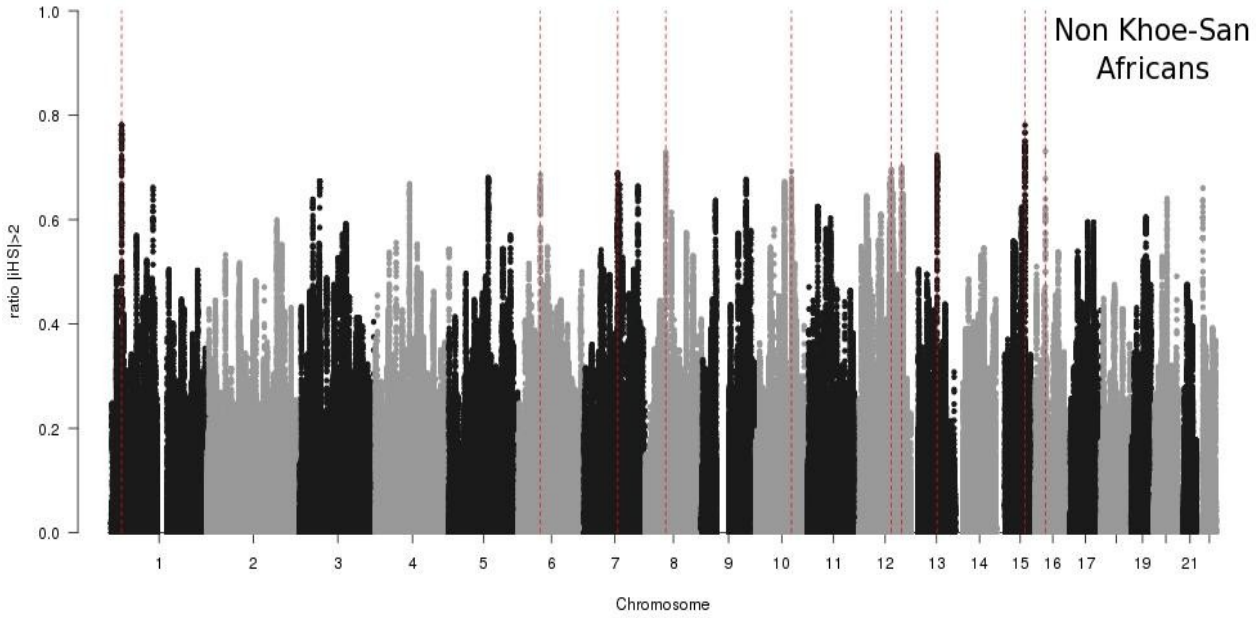


Figure S11.3: Manhattan plot of ratio |iHS|>2 in non Khoe-San Africans. Red dotted lines mark the position of the position of the highest |iHS| value in the ten candidate regions.

12. Selection in pre-modern humans

We relied on two different approaches to search for selection predating the Khoe-San split. One based on the 3P-CLR statistic (Racimo 2016) and one based on the PBS-framework used previously in (Schlebusch et al. 2012). Based on the genes in the top 10 regions, these two approaches were augmented with a GO-term enrichment analysis using DAVID (Huang, Brad T. Sherman, et al. 2009; Huang, Brad T Sherman, et al. 2009). The functions of the genes were retrieved from genecard.org (Stelzer et al. 2016), OMIM (Medicine, McKusick-Nathans Institute of Genetic, Johns Hopkins University (Baltimore) and UCSC Genome Browser (malacards (Rappaport et al. 2017) and uniprotkb (The UniProt Consortium 2019)).

12.1. 3P-CLR

We performed a selection scan using the 3P-CLR statistic (Racimo 2016) (Figure S12.1). This method is based on the XP-CLR statistic by (Nielsen et al. 2005) but contrasts three population instead of two and can specifically detect selection events on the internal branch. Here we targeted selection on the internal branch between the Neandertal/Denisovan – Homo sapiens split and the split between Khoe-San and other human populations. We used the two archaic individuals for the outgroup population, all western and eastern African populations in our global dataset and all KSP-individuals except the Nama from our data set together with the San from the HGDP-panel. We avoided the non-African populations due to their documented archaic admixture and the Nama were not included due to their eastern-African/Middle-eastern component. In total, the samples consisted of two archaic individuals, 22 eastern and western African individuals and 21 Khoe-San individuals.

The program calculates scores at every 20th SNP for which the outgroup genomes (Neandertal and/ or Denisovan) is polymorphic. 100 such SNPs were sampled in windows of size 0.25 cM. Below we describe the top windows for selection at the base of anatomically modern humans based on the maximum value of any SNP contained within the window:

Peak 1 (chr14:73685600-74304200, max value at 74156500): The SNP with the highest values is located inside the gene *DNAL1*. The encoded protein of this gene is expressed in tissues with motile cilia or flagella and may be involved in the movement of sperm flagella.

Peak 2 (chr6:117018000-117611000, max value at 117472000): The SNP with the highest score is located in an intergenic region downstreams of *RFX6* (insulin, diabetes, Mitchell-Riley syndrome) and upstreams of *VGLL2* (skeletal muscle development) and just after that downstreams of *ROS1* (a proto-oncogene that may function as a growth or differentiation factor receptor).

Peak 3 (chr1:41370500-41903800, max value at 41736600): Five SNPs (including the SNP with the highest score) with high scores are located within the gene *CTPS1* that codes for CTP Synthase 1. The activity of this protein is important for the immune system, and loss of function of this gene has been associated with immunodeficiency “malacards”: Immunodeficiency 24, also known as imd24, that has symptoms including lymphopenia and severe viral infections. Affiliated tissues include B-cells and T-cells. IMD24 is an autosomal recessive immunodeficiency characterized by the impaired capacity of activated T and B cells to proliferate in response to antigen receptor-mediated activation. Patients have early onset of severe chronic viral infections, mostly caused by herpesviruses, including Epstein-Barr virus (EBV) and varicella zoster virus (VZV); they also suffer from recurrent encapsulated bacterial infections, a spectrum typical of a combined deficiency of adaptive immunity (CID) (summary by (Kircher et al. 2014). Varicella zoster virus or varicella-zoster virus (VZV) is one of eight herpesviruses known to infect humans. VZV is a worldwide pathogen known by many names: chickenpox virus, varicella virus, zoster virus, and human herpesvirus type 3 (HHV-3). VZV infections are species-specific to humans, but can survive in external environments for a few hours, maybe a day or two. Epstein-Barr-virus or human herpesvirus 4 (HHV-4), is another of the eight known human herpesvirus types in the herpes family, and is one of the most common viruses in humans. The majority of all people are infected with EBV and gain adaptive immunity. It infects B cells of the immune system and epithelial cells.

Peak 4 (chr4:153058000-153590000, max value at 153297000): *FBXW7* (F-Box And WD Repeat Domain Containing 7, E3 Ubiquitin Protein Ligase) is a clear candidate gene in this region. Nine SNPs (including the SNP with the highest score) are located within *FBXW7*. The function of this gene is not well characterized. However, diseases associated with *FBXW7* include pediatric ependymoma and colon adenoma and among its related pathways are Immune System and Signaling by GPCR. It is involved in bone homeostasis and negative regulation of osteoclast differentiation (Fukushima et al. 2017).

Peak 5 (chr16:21176900-22209300, max value at 21649600): Three SNPs with more or less equally elevated scores covering a region with around ten genes at the beginning of chromosomal band 16p12.2. This is a structurally complex region susceptible to deletions and rearrangements. According to OMIM (Johns Hopkins University, Baltimore): The chromosome 16p12.2-p11.2 deletion syndrome is characterized phenotypically by dysmorphic facial features including flat faces, microretrognathia, blepharophimosis, short nose with hypoplastic nasal alae and absent nasal bridge, low-set and malformed ears, coloboma, and unilateral chorioretinitis. Other features included tetralogy of Fallot with pulmonary atresia, cubital deviation of the hands, talipes varus, articular limitation, and hemivertebra at level L1 as well as unilateral renal agenesis and cryptorchidism, feeding difficulties, recurrent ear infections, developmental delay, and cognitive impairment. Additional features, such as heart defects and short stature, are variable (Ballif et al. 2007; Battaglia et al. 2009). A 530 kb deletion (chr16:21.85-22.37 Mb) overlapping the peak region has been associated with developmental delay and craniofacial dysmorphism.

Peak 6 (chr4:20185400-20536700, max value at 20388600): All SNPs with elevated scores are located within the gene *SLIT2*. This gene encodes a member of the slit family of secreted glycoproteins, which are ligands for the Robo family of immunoglobulin receptors. Slit proteins play highly conserved roles in axon guidance and neuronal migration and may also have functions during other cell migration processes including leukocyte migration. SLIT1 and SLIT2 seem to be essential for midline guidance in the forebrain by acting as repulsive signal preventing inappropriate midline crossing by axons projecting from the olfactory bulb. In the developing visual system appears to function as repellent for retinal ganglion axons by providing a repulsion that directs these axons along their appropriate paths prior to, and after passage through, the optic chiasm.

Peak 7 (chr5:74279100-75132900, max value at 74700000): Many SNPs including the SNP with the highest score in this region overlaps the gene *COL4A3B4* (Homo sapiens collagen, type IV, alpha 3 (Goodpasture antigen) binding protein). Diseases associated with *COL4A3BP* include mental retardation, autosomal dominant 34 and goodpasture syndrome (a rare autoimmune disease in which antibodies attack the basement membrane in lungs and kidneys, leading to bleeding from the lungs and kidney failure). Among its related pathways are Metabolism and Sphingolipid metabolism (Sphingolipidoses, or disorders of sphingolipid metabolism, have particular impact on neural tissue).

Peak 8 (chr3:157186000-157930000, max value at 157411000): SNPs with elevated scores are located in between and downstream of C3orf55 and *SHOX2*. C3orf55 is an uncharacterized non-coding RNA with positive disease associations with alkaline phosphatase, body weight, iron, mental competency and vitamin K. *SHOX2* is a pseudoautosomal homeobox gene that is thought to be responsible for idiopathic short stature, and it is implicated in the short stature phenotype of Turner syndrome patients. This gene is considered to be a candidate gene for Cornelia de Lange syndrome. Most people with Cornelia de Lange syndrome have distinctive facial features, including arched eyebrows that often meet in the middle (synophrys), long eyelashes, low-set ears, small and widely spaced teeth, and a small and upturned nose. Many affected individuals also have behavior problems similar to autism, a developmental condition that affects communication and social interaction. Additional signs and symptoms of Cornelia de Lange syndrome can include excessive body hair (hypertrichosis), an unusually small head (microcephaly), hearing loss, and problems with the digestive tract.

The SNPs with elevated scores are closer to C3orf55 and are upstreams of *VEPH* (Homo sapiens ventricular zone expressed PH domain-containing 1) - a protein coding gene just just next to C3orf55. This gene is poorly characterized but its Drosophila homologue is a moderator of insulin (Teleman et al. 2005).

A known 8.9 Mb deletion including this region is located in (3q25.32) has been shown to be associated with facial dysmorphism with a coarse face, ptosis, synophrys, epicanthic folds, broad nasal bridge, long philtrum, large mouth with full lips, dysplastic and low-set ears (Moortgat et al. 2011).

Peak 9 (chr17:2984820-3437770, max value at 3076070): There are many SNPs with elevated 3P-CLR scores and they are all located in a region rich with olfactory receptors. Incidentally, the region is at the end of the (rather long) genomic band 17p13.3 and CNVs in this region are associated with Lissencephaly 1 the Miller–Dieker syndrome (genes implicated include PAFAH1B1 not so distant from the peak region). Lissencephaly 1 is characterized by smooth or nearly smooth cerebral surface and a paucity of gyral and sulcal development while the Miller–Dieker syndrome with Lissencephaly 1 as well as distinct characteristics including a short nose with upturned nares, thickened upper lip with a thin vermilion upper border, frontal bossing, small jaw, low-set posteriorly rotated ears, sunken appearance in the middle of the face, widely spaced eyes, and hypertelorism. The forehead is prominent with bitemporal hollowing. A microduplication on 17p13.3 (“17p13.3 microduplication syndrome”, ORPHA217385) is associated with discrete craniofacial dysmorphic features including a high forehead with frontal bossing, a small nose and a small mouth.

Peak 10 (chr4:128183000-129520000, max value at 129309000): SNPs with elevated score are spread out over a region with many genes. The SNP with the highest 3P-CLR score is located just upstreams of *PGRMC2* (progesterone receptor membrane component 2). Progesterone is an important sex hormone for both sexes and also functions as a neurosteroid.

Enrichment analysis for 3P-CLR using DAVID

As seen in Table S12.1, many GO-terms associated with neuron development seem to be enriched.

Summary

In summary, 3P-CLR provides support for that sexual selection may have been a significant selective force in the time period preceding the deepest divergence among modern humans (e.g. (Schlebusch et al. 2012)): sperm function is implicated in the top peak and peak 10 is linked to progesterone which is an important sex hormone. More incidental are that many peaks can be linked to physical appearance and as such a target for sexual selection: three (peak 5,8 and 9) out of the ten peaks can be linked to facial features of which peak 5 and 8 are also associated to stature and the peak 2 has a link to skeletal muscle development. Traces of pathogen driven selection can also be gleaned (peak 3,4 and 6) and neural development being the target of selection finds support in peak 5, 6, 7, 8 and 9. Selection on neural development is also supported by the GO-analysis. It is of course possible that sexual selection acts on neural development.

12.2. PBS-based statistics

Similar to (Schlebusch et al. 2012), we performed selection scans based on the PBS. PBS builds on transforming pairwise F_{ST} values between three populations into branch lengths in terms of drift. For an unrooted comparison with three populations there are three branch lengths and how the estimated branch lengths vary across genomic windows can be used to detect population specific selection. In (Schlebusch et al. 2012) a method that takes one of the branches to be the outgroup and then search for regions with an unusually long estimated outgroup branch was utilized to detect selection prior to, but close to, the split between the Khoe-San populations and other human populations. This statistic was referred to as the aPBS.

Here we included two archaic human genomes (Neandertal and Denisovan) in the “KSP+HGDP” group-called dataset. This allowed us to set up two additional statistics similar to the aPBS that

would also target selection close to but prior to the Khoe-San split. Instead of the great apes as an outgroup population we used the two archaic genomes as the outgroup population but conditioned on variants being polymorphic in the combined two modern human populations. We refer to this statistic as archaicPBS. Similar to how pairwise F_{ST}-values can be combined to estimate specific branch lengths for three populations, it is also possible to combine pairwise F_{ST}-values in a four-population set-up. This allows us to solve for the branch length of the internal branch (there is exactly one internal branch in an unrooted 4-population set-up). More specifically, denote the four populations are denoted by *A*, *B*, *C* and *D*, the external branches by *a*, *b*, *c* and *d* and the internal branch by *i* (see Figure S12.2). If it is known that A and B are closer related to each other than to any other population, the length of the internal branch (*i*) can be estimated as

$[f(A,D) + f(B,D) + f(A,C) + f(B,C) - 2f(A,B) - 2f(C,D)]/4$

$=[(a+i+d) + (b+i+d) + (a+i+c) + (b+i+c) + (a+i+d) - 2(a+b) - 2(c+d)]/4$

$= i$

where $f(x,y)=1-\ln(F_{ST}(x,y))$ and F_{ST} is calculated following (Weir and Cockerham 1984).

Figure S12.2 also shows for what time periods in the human phylogeny the three statistics are expected to pick up selection signals.

Using the two human populations for A and B, the two archaic genomes for C and the outgroup for D, we refer to this statistic as emhPBS.

For the aPBS, archaicPBS and emhPBS selection scans (Figure S12.3), we relied on the "KSP+HGDP" group-called dataset as well as the two archaic individuals. In order to avoid (as much as possible) biases due to different sample sizes, we picked two individuals from the Khoe-San branch (an individual from the Ju|'hoansi population in our sample and the San individual (also Ju|'hoansi) from the HGDP dataset), two individuals from the non-Khoe-San branch (the Yoruba and Mandenka individuals from the HGDP data set) as well as the two archaic individuals. The genome was divided into 100 kb windows with 1 kb steps in between. After calculating the statistics for each of the (included) windows, overlapping windows in the top 1% were merged into regions. Each of these regions were given a value --a height-- as the maximum among its windows and a width --the number of windows that the region was constructed from. Similar to (Schlebusch et al. 2012) we searched for peaks that were unusually wide and/or unusually high. For this purpose, among the regions constructed from the top 1 % windows (before merging) we calculated a distance for each region that should weigh width and height equally by first calculating for each region in our list:

x*=height - minimum height among studied regions

y*=width - minimum width among studied regions

then

x=x*/max(x*)

y=y*/max(y*)

to finally get the (euclidean) distance

$d=\mathrm{sqrt}(x^2+y^2)$

This distance should give equal weight to height and width and we choose to study the top five regions according to this d-distance. They are listed in Table S12.2.

We also calculated the PBS values for the Khoe-San branch (the two Ju|'hoansi individuals, denoted by KSP-PBS) and for the non-Khoe-San branch (the Yoruba and the Mandenka individuals, denoted by nonKSP-PBS). These statistics were not optimized to detect branch specific selection as the

sample size is very small but may still serve as a background that the three PBS statistics devised to detect ancient selection can be compared to.

In general aPBS, archaicPBS and emhPBS were strongly correlated and one region show up in the top five regions for all three statistics. In fact, studying the top ten regions across the three statistics show that five regions show up in all three top ten lists and in all but one case, all three statistics show elevated values even if they are not within the top ten regions.

Region 1) The region that shows up in the top five lists for all three PBS-statistics devised to detect ancient selection also shows up in the top five regions for KSP-PBS. There is an obvious effect also on the nonKSP-PBS statistic. This region is located on chromosome 4 from 61096000 to 61946000. Although no gene is located within this region, it is just upstreams of the gene *LPHN3* (starts at around 62.07 Mb). *LPHN3* encodes a member of the latrophilin subfamily of G-protein coupled receptors (GPCR). Diseases associated with this gene include attention deficit-hyperactivity disorder and it is important for determining the connectivity rates between the principal neurons in the cortex.

Region 2) Not far away from this region, the region chr4:86.1-86.3 Mb is present in the top ten lists of aPBS, archaicPBS and emhPBS and on the top five lists for aPBS and archaicPBS. Here, there is also a slight effect on nonKSP-PBS. There is no coding region within this region but it is located upstreams of *ARHGAP24* (Homo sapiens Rho GTPase activating protein 24, starts at 86.4). ARHGAPs, such as *ARHGAP24*, encode negative regulators of Rho GTPases, which are implicated in actin remodeling, cell polarity, and cell migration. Diseases associated with *ARHGAP24* include familial idiopathic steroid-resistant nephrotic syndrome with focal segmental glomerulosclerosis (related to the function of the kidney) and atypical autism.

Region 3) A peak that shows up on the top five list for aPBS and emhPBS but not for archaicPBS is the region chr4:70116000-70232000. This is the region with the highest aPBS across the genome but also emhPBS and archaicPBS are elevated here. The genomic region with elevated values is clearly defined and is just downstreams of *UGT2B28* (Homo sapiens UDP glucuronosyltransferase 2 family, polypeptide B28 chr4: 70 146 217-70 160 768). This gene encodes an enzyme that catalyzes the transfer of glucuronic acid from uridine diphosphoglucuronic acid to a diverse array of substrates including steroid hormones and lipid-soluble drugs. This process, known as glucuronidation, is an intermediate step in the metabolism of steroids. Among its related pathways are Metabolism and Chemical carcinogenesis. UDPGTs are of major importance in the conjugation and subsequent elimination of potentially toxic xenobiotics and endogenous compounds. The neighboring genes to *UGT2B28* are also UGT2B-genes (*UGT2B11* and *UGT2B4*).

Region 4) The fourth peak for aPBS is also on the top five for archaicPBS but not even on the top ten list for emhPBS. It is however also present on the top five for KSP-PBS. In fact, all five PBS statistics are elevated in this region but no other statistic. The region (chr7:106679000-107270000, elevated values across the whole region) is well covered by the gene *COG5* (Homo sapiens component of oligomeric golgi complex 5). Mutations in this gene result in congenital disorder of glycosylation type 2I (malacards: Congenital disorders of glycosylation result in a wide variety of clinical features, such as defects in the nervous system development, psychomotor retardation, dysmorphic features, hypotonia, coagulation disorders, and immunodeficiency). *GPR22* (Homo sapiens G protein-coupled receptor 22) is another interesting (short) gene located within the region as well as within the boundaries for *COG5*. This gene is expressed in the brain regions frontal cortex, caudate, putamen and thalamus; not in pons, hypothalamus and hippocampus. It is also a region just upstream of *BCAP29* which is a B-cell receptor and thus potentially important for the immune system.

Region 5) The fifth peak (chr17:43660000-44465000) for aPBS is on no other list. There is no visible effect on any other PBS statistic. The region covers three major genes CRHR1, MAPT and KANSL1. It is also a region polymorphic for the microtubular associated protein tau (MAPT) inversion, which is an ~900 kb inversion (see (Donnelly et al. 2010)). Whether the signal in aPBS is an artifact of this inversion or not is unclear at this point but this is an interesting region with

strong connections to brain function characterized by moderate to severe intellectual disability, hypotonia, friendly demeanor, and highly distinctive facial features, including tall, broad forehead, long face, upslanting palpebral fissures, epicanthal folds, tubular nose with bulbous nasal tip, and large ears.

Region 6) The region chr2:63.1-63.6 Mb is on the top ten lists for all three statistics and on the top five lists for archaicPBS and emhPBS. The gene *WDPCP* (Homo sapiens WD repeat containing planar cell polarity effector) is implicated here. A similar gene in frogs encodes a planar cell polarity protein that plays a critical role in collective cell movement and ciliogenesis by mediating septin localization. Mutations in this gene are associated with Bardet-Biedl syndrome 15 and may also play a role in Meckel-Gruber syndrome.

Region 7) Similar to the previous region, chr16:47077000-47821000 (elevated values suggest 47.3 to 47.8 Mb) is on the top ten lists for all three statistics and on the top five lists for archaicPBS and emhPBS. The statistics aPBS, archaicPBS and emhPBS show quite evenly elevated values across this genomic region suggesting one of two genes *ITFG1* (Homo sapiens integrin alpha FG-GAP repeat containing 1) or *PHKB* (Homo sapiens phosphorylase kinase, beta). *ITFG1* is a modulator of T-cell function and mutations in *PHKB* cause glycogen storage disease type 9B, also known as phosphorylase kinase deficiency of liver and muscle.

Region 8) The only peak in the top five list for emhPBS not on either of the top five lists for aPBS and/or archaicPBS is the region chr3:93453000-93982000. Also aPBS is affected but not archaicPBS. Interestingly, there is also a signal for 3P-CLR-anc and XP-EHH KSP-vs-nonKSP is clearly positive while XP-EHH N-KSP-vs-S-KSP is clearly negative suggesting selection in the Nama and Karretjie. The region is quite close to the centromere. Candidate genes include *PROS1* (Homo sapiens protein S (alpha) – encodes a protein that functions as a cofactor for the anticoagulant), *ARL13B* (Homo sapiens ADP-ribosylation factor-like 13B – involved in cerebral cortex development and mutations in this gene are associated with Joubert syndrome 8), *STX19* (Homo sapiens syntaxin 19 – among its related pathways is the Nicotine Pathway) and *DHFRL1* (Homo sapiens dihydrofolate reductase-like 1 – a key enzyme in folate metabolism).

Enrichment analysis for three PBS-statistics using DAVID

As evident in Tables S12.3 to S12.5, GO-terms associated with flagellum development are especially enriched for emhPBS but also for aPBS. The number of enriched terms is clearly lower than for 3P-CLR and for archaicPBS there is a single term. The enrichment of GO-terms associated with neural development found for 3P-CLR is not evident for these statistics (except for one term in emhPBS).

Summary

Similar to the 3P-CLR analysis, brain development (regions 1,4,5,8), bodily defense systems (regions 3 and 7) and sexual selection/sperm motility (GO-analysis) appears to be a red thread also among the candidate genes based on the three PBS statistics.

Selection in pre-modern humans Tables and Figures

Table S12.1: GO-terms with FDR<10% among 224 genes in top regions for 3P-CLR.

Term	Count	%	PValue	Fold Enrichment	FDR
GO:0016337~cell-cell adhesion	15	6.52	2.21E-05	4.01	0.0368
GO:0007155~cell adhesion	24	10.43	5.24E-05	2.57	0.0870
GO:0022610~biological adhesion	24	10.43	5.36E-05	2.56	0.0891
GO:0048666~neuron development	15	6.52	0.000178	3.30	0.295
GO:0048812~neuron projection morphogenesis	11	4.78	0.000508	3.91	0.841
GO:0032990~cell part morphogenesis	12	5.22	0.000592	3.53	0.979
GO:0000902~cell morphogenesis	14	6.09	0.000867	2.97	1.43
GO:0007409~axonogenesis	10	4.35	0.00109	3.89	1.79
GO:0048858~cell projection morphogenesis	11	4.78	0.00150	3.39	2.47
GO:0048667~cell morphogenesis involved in neuron differentiation	10	4.35	0.00182	3.61	2.97
GO:0031175~neuron projection development	11	4.78	0.00207	3.25	3.38
GO:0040017~positive regulation of locomotion	7	3.04	0.00219	5.21	3.58
GO:0006464~protein modification process	33	14.35	0.00230	1.72	3.75
GO:0032989~cellular component morphogenesis	14	6.09	0.00230	2.66	3.75
GO:0030182~neuron differentiation	15	6.52	0.00235	2.54	3.84
GO:0007275~multicellular organismal development	55	23.91	0.00261	1.45	4.24
GO:0048468~cell development	19	8.26	0.00264	2.18	4.29
GO:0032502~developmental process	59	25.65	0.00278	1.41	4.52
GO:0043687~post-translational protein modification	28	12.17	0.00323	1.79	5.23
GO:0022008~neurogenesis	18	7.83	0.00339	2.19	5.49
GO:0048699~generation of neurons	17	7.39	0.00399	2.22	6.42
GO:0009896~positive regulation of catabolic process	5	2.17	0.00437	7.45	7.01
GO:0030163~protein catabolic process	17	7.39	0.00484	2.18	7.74
GO:0043412~biopolymer modification	33	14.35	0.00487	1.64	7.79
GO:0000904~cell morphogenesis involved in differentiation	10	4.35	0.00500	3.10	7.99

Table S12.2: Regions in top five PBS according to d-distance (see text).

aPBS

	chr	startpos	endpos	maxpos	maxval	numWindows
	2	63094000	63609000	63419000	1.99104	388
	3	35253000	35759000	35535000	2.384943	392
peak1	4	61169000	61884000	61307000	3.041188	611
peak2	4	70116000	70232000	70182000	3.823291	17
peak3	4	86040000	86344000	86280000	2.928521	205
	7	83626000	83789000	83699000	2.814628	64
peak4	7	106679000	107270000	107082000	2.40519	489
	16	14578000	15053000	14815000	2.709499	245
	16	47233000	47805000	47356000	2.208242	426
peak5	17	43660000	44465000	44392000	2.59391	706

archaicPBS

	chr	startpos	endpos	maxpos	maxval	numWins
peak1	2	63047000	63638000	63530000	2.221352	492
	3	89337000	89709000	89475000	2.424003	216
peak2	4	61192000	61883000	61359000	2.759333	590
peak3	4	85930000	86344000	86159000	2.639083	216
peak4	7	106678000	107292000	107082000	2.550082	515
	10	74917000	75411000	75090000	2.032877	337
	14	71843000	72247000	72094000	2.132352	303
	16	14550000	14873000	14761000	2.443669	224
peak5	16	47077000	47821000	47340000	2.283413	584
	18	46638000	47027000	46937000	2.219182	289

PBS_earlyModernHuman

	chr	startpos	endpos	maxpos	maxval	numWins
	1	103906000	104254000	103957000	1.476617	238
peak1	2	62985000	63609000	63419000	1.129241	507
peak2	3	93453000	93982000	93848000	1.308769	415
	4	9147000	9269000	9210000	1.535563	23

peak3	4	61194000	61724000	61343000	1.381818	392
peak4	4	70116000	70232000	70180000	1.939314	17
	4	85470000	86000000	85742000	1.104478	383
	14	71804000	72253000	72093000	1.361253	327
	16	14498000	14873000	14815000	1.50841	264
peak5	16	47068000	47832000	47340000	1.44766	621

Table S12.3: GO-terms with FDR<10% among 224 genes in top regions for aPBS**.**

Term	Count	%	PValue	Fold Enrichment	FDR
GO:0019932~second-messenger-mediated signaling	12	4.40	0.00235	2.980	3.80
GO:0044460~flagellum part	3	1.10	0.00431	28.65	5.63
GO:0044442~microtubule-based flagellum part	3	1.10	0.00431	28.65	5.63
GO:0005887~integral to plasma membrane	32	11.72	0.00556	1.66	7.21
GO:0031226~intrinsic to plasma membrane	32	11.72	0.00772	1.62	9.87

Table S12.4: GO-terms with FDR<10% among 224 genes in top regions for archaicPBS**.**

Term	Count	%	PValue	Fold Enrichment	FDR
GO:0019932~second-messenger-mediated signaling	12	4.11	0.00400	2.78	6.49

Table S12.5: GO-terms with FDR<10% among 224 genes in top regions for emhPBS**.**

Term	Count	%	PValue	Fold Enrichment	FDR
GO:0030030~cell projection organization	15	5.10	0.00473	2.36	7.63
GO:0044460~flagellum part	3	1.02	0.00489	26.87	6.38
GO:0044442~microtubule-based flagellum part	3	1.02	0.00489	26.87	6.38
GO:0009434~microtubule-based flagellum	4	1.36	0.00490	11.31	6.39
GO:0031175~neuron projection development	12	4.08	0.00506	2.70	8.13
GO:0042221~response to chemical stimulus	36	12.24	0.00567	1.59	9.06
GO:0019861~flagellum	5	1.70	0.00624	6.72	8.08

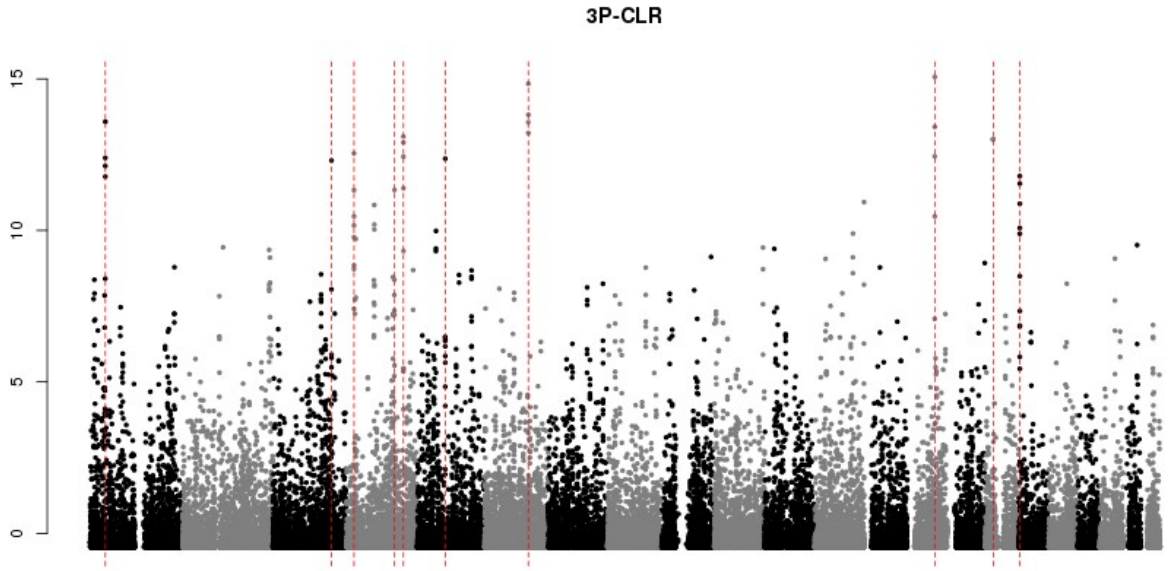


Figure S12.1: Manhattan plot of 3P-CLR (Racimo 2016) with Neandertal+Denisovan versus the clade Khoe-San/non-Khoe San Africans.

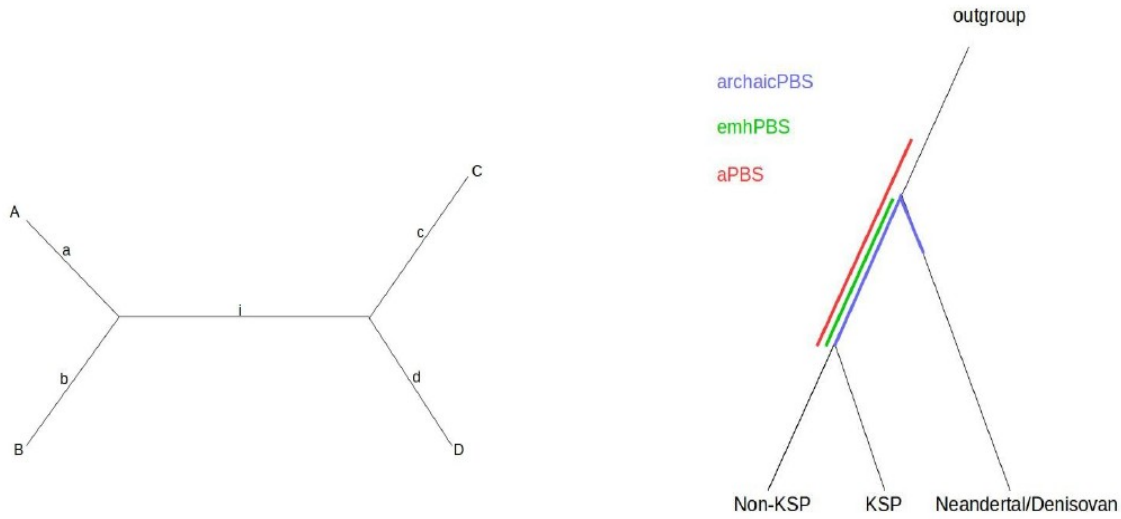


Figure S12.2: Left panel shows the notation for calculating emhPBS. Right panel illustrates the difference in terms of where in the phylogeny the three statistics archaicPBS, emhPBS and aPBS are potentially sensitive to selective sweeps.

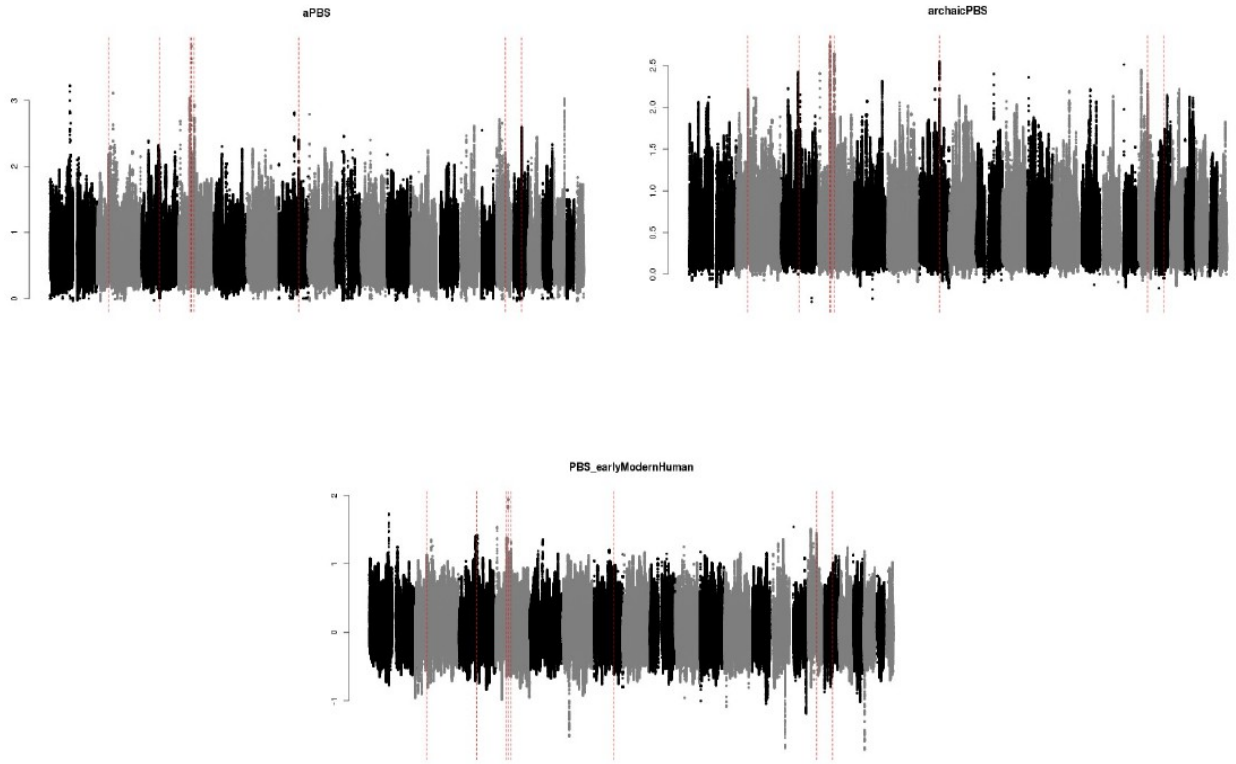


Figure S12.3: Manhattan plot of the three PBS-derived statistics used to detect selection prior to the deepest split among modern human populations. Red horizontal lines show the positions of the top peaks.

Literature cited

Abeler-Dörner L, Swamy M, Williams G, Hayday AC, Bas A. 2012. Butyrophilins: an emerging family of immune regulators. Trends Immunol. [Internet] 33:34–41. Available from: http://www.sciencedirect.com/science/article/pii/S1471490611001633

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Mol. Biol. Evol. [Internet] 27:2534–2547. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msq148

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. [Internet] 19:1655–1664. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19648217

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. Nature [Internet] 526:68–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26432245

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinforma. [Internet] 43:1–33. Available from: http://doi.wiley.com/10.1002/0471250953.bi1110s43

Ballif BC, Hornor SA, Jenkins E, Madan-Khetarpal S, Surti U, Jackson KE, Asamoah A, Brock PL, Gowans GC, Conway RL, et al. 2007. Discovery of a previously unrecognized microdeletion syndrome of 16p11.2–p12.2. Nat. Genet. [Internet] 39:1071–1073. Available from: http://www.nature.com/doifinder/10.1038/ng2107

Barbieri C, Güldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, Mpoloka SW, Stoneking M, Pakendorf B. 2014. Unraveling the complex maternal history of Southern African Khoisan populations. Am. J. Phys. Anthropol. [Internet] 153:435–448. Available from: http://doi.wiley.com/10.1002/ajpa.22441

Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. 2013. Ancient Substructure in Early mtDNA Lineages of Southern Africa. Am. J. Hum. Genet. [Internet] 92:285–292. Available from: https://www.sciencedirect.com/science/article/pii/S0002929712006441?via%3Dihub

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F. 2007. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol. Phylogenet. Evol. [Internet] 43:635–644. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1055790306003708

Battaglia A, Novelli A, Bernardini L, Igliozzi R, Parrini B. 2009. Further characterization of the new microdeletion syndrome of 16p11.2-p12.2. Am. J. Med. Genet. Part A [Internet] 149A:1200–1204. Available from: http://doi.wiley.com/10.1002/ajmg.a.32847

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, et al. 2008. The dawn of human matrilineal diversity. Am. J. Hum. Genet. [Internet] 82:1130–1140. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929708002553

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics [Internet] 32:2817–2823. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw327

Beichman AC, Phung TN, Lohmueller KE. 2017. Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. G3 (Bethesda). [Internet] 7:3605–3620. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28893846

Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. 2009. Next generation software for functional trend analysis. Bioinformatics [Internet] 25:3043–3044. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp498

Birtle Z, Goodstadt L, Ponting C. 2005. Duplication and positive selection among hominin-specific PRAME genes. BMC Genomics [Internet] 6:120. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-6-120

Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. 2014. Lactase persistence alleles reveal partial east African ancestry of southern African Khoe pastoralists. Curr. Biol. 24.

Broad Institute. Picard. Available from: http://broadinstitute.github.io/picard/

Browning SR, Browning BL. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. [Internet] 81:1084–1097. Available from: http://www.sciencedirect.com/science/article/pii/S0002929707638828?via%3Dihub

Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, Amenga-Etego LN, Enimil A, Apinjoh T, Ndila CM, et al. 2016. Admixture into and within sub-Saharan Africa. Elife [Internet] 5. Available from: https://elifesciences.org/articles/15266

Busing FMTA, Meijer E, Leeden R Van Der. 1999. Delete-m Jackknife for Unequal m. Stat. Comput. [Internet] 9:3–8. Available from: http://link.springer.com/10.1023/A:1008800423698

Carty CL, Johnson NA, Hutter CM, Reiner AP, Peters U, Tang H, Kooperberg C. 2012. Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). Hum. Mol. Genet. [Internet] 21:711–720. Available from: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddr489

Cassidy SB. 1997. Prader-Willi syndrome. J. Med. Genet. [Internet] 34:917–923. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9391886

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. Genome Res. [Internet] 19:136–142. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19029539

Cheng A, Solomon MJ. 2008. Speedy/Ringo C regulates S and G $_2$ phase progression in human cells. Cell Cycle [Internet] 7:3037–3047. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18802405

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:

SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). [Internet] 6:80–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22728672

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics [Internet] 27:2156–2158. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods [Internet] 9:772. Available from: http://www.nature.com/articles/nmeth.2109

Dastani Z, Hivert M-F, Timpson N, Perry JRB, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikäinen L-P, et al. 2012. Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals.Visscher PM, editor. PLoS Genet. [Internet] 8:e1002607. Available from: http://dx.plos.org/10.1371/journal.pgen.1002607

Delaneau O, Marchini J, Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. Nat. Methods [Internet] 9:179–181. Available from: http://www.nature.com/articles/nmeth.1785

DeLorey TM, Handforth A, Anagnostaras SG, Homanics GE, Minassian BA, Asatourian A, Fanselow MS, Delgado-Escueta A, Ellison GD, Olsen RW. 1998. Mice lacking the beta3 subunit of the GABAA receptor have the epilepsy phenotype and many of the behavioral characteristics of Angelman syndrome. J. Neurosci. [Internet] 18:8505–8514. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9763493

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. [Internet] 43:491–498. Available from: http://www.nature.com/articles/ng.806

Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Mehdi SQ, Kajuna SLB, Barta C, Kungulilo S, Karoma NJ, Lu R-B, et al. 2010. The distribution and most recent common ancestor of the 17q21 inversion in humans. Am. J. Hum. Genet. [Internet] 86:161–171. Available from: http://linkinghub.elsevier.com/retrieve/pii/S000292971000008X

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science [Internet] 327:78–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19892942

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Mol. Biol. Evol. [Internet] 22:1185–1192. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15703244

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. [Internet] 29:1969–1973. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss075

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. 2016. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. Mol. Biol. Evol. [Internet] 33:1082–1093. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26715629

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. Nucleic Acids Res. [Internet] 41:D48-55. Available from: http://academic.oup.com/nar/article/41/D1/D48/1070752/Ensembl-2013

Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. 2017. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. [Internet] 45:D777–D783. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1121

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, et al. 2012. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature [Internet] 493:216–220. Available from: http://www.nature.com/doifinder/10.1038/nature11690

Fukushima H, Shimizu K, Watahiki A, Hoshikawa S, Kosho T, Oba D, Sakano S, Arakaki M, Yamada A, Nagashima K, et al. 2017. NOTCH2 Hajdu-Cheney Mutations Escape SCFFBW7-Dependent Proteolysis to Promote Osteoporosis. Mol. Cell [Internet] 68:645-658.e5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29149593

Geetha C, Venkatesh SG, Dunn BHF, Gorr S-U. 2003. Expression and anti-bacterial activity of human parotid secretory protein (PSP). Biochem. Soc. Trans. [Internet] 31:815–818. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12887312

Gessain A, Cassar O. 2012. Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front. Microbiol. [Internet] 3:388. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23162541

Van Geystelen A, Decorte R, Larmuseau MH. 2013. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics [Internet] 14:101. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23405914

Gilad Y, Bustamante CD, Lancet D, Pääbo S. 2003. Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am. J. Hum. Genet. [Internet] 73:489–501. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12908129

Goubau P, Desmyter J, Swanson P, Reynders M, Shih J, Surmont I, Kazadi K, Lee H. 1993. Detection of HTLV-I and HTLV-II infection in Africans using type-specific envelope peptides. J. Med. Virol. [Internet] 39:28–32. Available from: http://doi.wiley.com/10.1002/jmv.1890390107

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. Science [Internet] 328:710–722. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20448178

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. [Internet] 43:1031–1034. Available from: http://www.nature.com/articles/ng.937

Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson B V, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, et al. 2008. Many sequence variants affecting diversity of adult human height. Nat. Genet. [Internet] 40:609–615. Available from: http://www.nature.com/doifinder/10.1038/ng.122

Hallast P, Rull K, Laan M. 2007. The evolution and genomic landscape of CGB1 and CGB2 genes. Mol. Cell. Endocrinol. [Internet] 260–262:2–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17055150

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. 2015. Large multiallelic copy number variations in humans. Nat. Genet. [Internet] 47:296–303. Available from: http://www.nature.com/doifinder/10.1038/ng.3200

Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Mägi R, et al. 2010. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat. Genet. [Internet] 42:949–960. Available from: http://www.nature.com/doifinder/10.1038/ng.685

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. U. S. A. [Internet] 108:5154–5162. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1017511108

Hilton HG, Norman PJ, Nemat-Gorgani N, Goyos A, Hollenbach JA, Henn BM, Gignoux CR, Guethlein LA, Parham P. 2015. Loss and Gain of Natural Killer Cell Receptor Function in an African Hunter-Gatherer Population.Tishkoff SA, editor. PLOS Genet. [Internet] 11:e1005439. Available from: http://dx.plos.org/10.1371/journal.pgen.1005439

Hollfelder N, Schlebusch CM, Günther T, Babiker H, Hassan HY, Jakobsson M. 2017. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations.Tishkoff SA, editor. PLOS Genet. [Internet] 13:e1006976. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28837655

Huang DW, Sherman Brad T., Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. [Internet] 37:1–13. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn923

Huang DW, Sherman Brad T, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. [Internet] 4:44–57. Available from: http://www.nature.com/articles/nprot.2008.211

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics [Internet] 18:337–338. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11847089

Johns Hopkins University, Baltimore M. Online Mendelian Inheritance in Man, OMIM®. MIM Number: 613604. Available from: https://omim.org/

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature [Internet] 413:514–519. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11586358

Kanwar YS, Kumar A, Yang Q, Tian Y, Wada J, Kashihara N, Wallner EI. 1999. Tubulointerstitial nephritis antigen: an extracellular matrix protein that selectively regulates tubulogenesis vs. glomerulogenesis during mammalian renal development. Proc. Natl. Acad. Sci. U. S. A. [Internet] 96:11323–11328. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10500175

Khan I, Maldonado E, Vasconcelos V, O'Brien SJ, Johnson WE, Antunes A. 2014. Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments. BMC Genomics [Internet] 15:779. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25208914

Kim SA, Kim JH, Park M, Cho IH, Yoo HJ. 2006. Association of GABRB3 Polymorphisms with Autism Spectrum Disorders in Korean Trios. Neuropsychobiology [Internet] 54:160–165. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17230033

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. [Internet] 46:310–315. Available from: http://www.nature.com/articles/ng.2892

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. 2004. Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. Am. J. Hum. Genet. [Internet] 75:752–770. Available from: https://www.sciencedirect.com/science/article/pii/S0002929707637835?via%3Dihub

Kofler R, Schlötterer C. 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. Bioinformatics [Internet] 28:2084–2085. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22635606

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir Aslaug, Walters GB, Jonasdottir Adalbjorg, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature [Internet] 467:1099–1103. Available from: http://www.nature.com/doifinder/10.1038/nature09525

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays [Internet] 35:780–786. Available from: http://doi.wiley.com/10.1002/bies.201300014

Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo J-M, Lema G, Fu W, Nyambo TB, Rebbeck TR, et al. 2012. Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. Cell [Internet] 150:457–469. Available from: http://www.sciencedirect.com/science/article/pii/S0092867412008318?via%3Dihub

Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in

60,706 humans. Nature [Internet] 536:285–291. Available from: http://www.nature.com/articles/nature19057

Lengyel P. 1982. Biochemistry of interferons and their actions. Annu. Rev. Biochem. [Internet] 51:251–282. Available from: http://www.annualreviews.org/doi/10.1146/annurev.bi.51.070182.001343

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet] 25:1754–1760. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19451168

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature [Internet] 475:493–496. Available from: http://www.nature.com/articles/nature10231

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet] 25:2078–2079. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943

Li S, Schlebusch C, Jakobsson M. 2014. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proceedings. Biol. Sci. [Internet] 281:20141448. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25209939

Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, Speliotes EK, Thorleifsson G, Willer CJ, Herrera BM, et al. 2009. Genome-Wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution.Allison DB, editor. PLoS Genet. [Internet] 5:e1000508. Available from: http://dx.plos.org/10.1371/journal.pgen.1000508

Lombard M, Parsons I. 2015. Milk not Meat: The Role of Milk amongst the Khoe Peoples of Southern Africa. J. African Archaeol. [Internet] 13:149–166. Available from: http://www.african-archaeology.de/index.php?page_id=154&journal_id=40&pdf_id=311

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science [Internet] 335:823–828. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22344438

MacArthur DG, Tyler-Smith C. 2010. Loss-of-function variants in the genomes of healthy humans. Hum. Mol. Genet. [Internet] 19:R125–R130. Available from: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddq365

Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, Pakendorf B, Stoneking M. 2014. Tracing Pastoralist Migrations to Southern Africa with Lactase Persistence Alleles. Curr. Biol. [Internet] 24:875–879. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24704073

Mahdi T, Hänzelmann S, Salehi A, Muhammed SJ, Reinbothe TM, Tang Y, Axelsson AS, Zhou Y, Jing X, Almgren P, et al. 2012. Secreted Frizzled-Related Protein 4 Reduces Insulin Secretion and Is Overexpressed in Type 2 Diabetes. Cell Metab. [Internet] 16:625–633. Available from: https://www.sciencedirect.com/science/article/pii/S1550413112004093

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature [Internet] 538:201–206. Available from: http://www.nature.com/articles/nature18964

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics [Internet] 26:2867–2873. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq559

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. [Internet] 93:278–288. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929713002899

Marshall L, Obaidullah M, Fuchs T, Fineberg NS, Brinkley G, Mikuls TR, Bridges SL, Hermel E, Hermel E. 2014. CASPASE-12 and rheumatoid arthritis in African-Americans. Immunogenetics [Internet] 66:281–285. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24515649

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier J-B, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med. [Internet] 6:26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24944579

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. [Internet] 20:1297–1303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20644199

McKie AT, Barrow D, Latunde-Dada GO, Rolfs A, Sager G, Mudaly E, Mudaly M, Richardson C, Barlow D, Bomford A, et al. 2001. An iron-regulated ferric reductase associated with the absorption of dietary iron. Science [Internet] 291:1755–1759. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11230685

Medicine, McKusick-Nathans Institute of Genetic, Johns Hopkins University (Baltimore M. Online Mendelian Inheritance in Man, OMIM®. Available from: https://omim.org/

Menashe I, Man O, Lancet D, Gilad Y. 2003. Different noses for different people. Nat. Genet. [Internet] 34:143–144. Available from: http://www.nature.com/articles/ng1160

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science (80-.). 338:222–226.

Mi H, Thomas P. 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol. Biol. [Internet] 563:123–140. Available from: http://link.springer.com/10.1007/978-1-60761-175-2_7

Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. [Internet] 16:1215. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3344216

Moortgat S, Verellen-Dumoulin C, Maystadt I, Parmentier B, Grisart B, Hennecker J-L, Destree A. 2011. Developmental delay and facial dysmorphism in a child with an 8.9 Mb de novo interstitial deletion of 3q25.1-q25.32: Genotype-phenotype correlations of chromosome 3q25 deletion syndrome. Eur. J. Med. Genet. [Internet] 54:177–180. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1769721210001436

N'Diaye A, Chen GK, Palmer CD, Ge B, Tayo B, Mathias RA, Ding J, Nalls MA, Adeyemo A, Adoue V, et al. 2011. Identification, Replication, and Fine-Mapping of Loci Associated with Adult Height in Individuals of African Ancestry.Visscher PM, editor. PLoS Genet. [Internet] 7:e1002298. Available from: http://dx.plos.org/10.1371/journal.pgen.1002298

Naidoo T, Sjödin P, Schlebusch C, Jakobsson M. 2018. Patterns of variation in cis-regulatory regions: examining evidence of purifying selection. BMC Genomics [Internet] 19:95. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4422-y

Naidoo T, Xu J, Vicente M, Malmström H, Soodyall H, Jakobsson M, Schlebusch C. In revision. Y chromosome variation in southern African Khoe-San populations based on whole genome sequences. Genome Biology and Evolution.

Need AC, Attix DK, McEvoy JM, Cirulli ET, Linney KL, Hunt P, Ge D, Heinzen EL, Maia JM, Shianna K V, et al. 2009. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. Hum. Mol. Genet. [Internet] 18:4650–4661. Available from: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddp413

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. Nat. Rev. Genet. [Internet] 8:857–868. Available from: http://www.nature.com/doifinder/10.1038/nrg2187

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. [Internet] 15:1566–1575. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16251466

Nusse R. 2005. Wnt signaling in disease and in development. Cell Res. [Internet] 15:28–32. Available from: http://www.nature.com/articles/7290260

Nuwal T, Kropp M, Wegener S, Racic S, Montalban I, Buchner E. 2012. The *Drosophila* Homologue of Tubulin-Specific Chaperone E–Like Protein Is Required for Synchronous Sperm Individualization and Normal Male Fertility. J. Neurogenet. [Internet] 26:374–381. Available from: http://www.tandfonline.com/doi/full/10.3109/01677063.2012.731119

Okonechnikov K, Conesa A, García-Alcalde F. 2015. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics [Internet] 32:btv566. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26428292

Oudhoff MJ, Bolscher JGM, Nazmi K, Kalay H, van 't Hof W, Amerongen AVN, Veerman ECI. 2008. Histatins are the major wound-closure stimulating factors in human saliva as identified in a cell culture assay. FASEB J. [Internet] 22:3805–3812. Available from: http://www.fasebj.org/doi/10.1096/fj.08-112003

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. [Internet] 30:E386-94. Available from: http://doi.wiley.com/10.1002/humu.20921

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, et al. 2012. Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. Am. J. Hum. Genet. [Internet] 91:83–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22726845

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert J-M, et al. 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set.Di Rienzo A, editor. PLoS Genet. [Internet] 5:e1000448. Available from: http://dx.plos.org/10.1371/journal.pgen.1000448

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science (80-.). [Internet] 356:543–546. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28473590

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. Genetics [Internet] 192:1065–1093. Available from: http://www.genetics.org/cgi/doi/10.1534/genetics.112.145037

Patterson N, Price AL, Reich D. 2006. Population Structure and Eigenanalysis. PLoS Genet. [Internet] 2:e190. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17194218

Peterson JA, Scallan CD, Ceriani RL, Hamosh M. 2001. Structural and functional aspects of three major glycoproteins of the human milk fat globule membrane. Adv. Exp. Med. Biol. [Internet] 501:179–187. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11787681

Phung TN, Huber CD, Lohmueller KE. 2016. Determining the Effect of Natural Selection on Linked Neutral Divergence across Species.Akey JM, editor. PLOS Genet. [Internet] 12:e1006199. Available from: http://dx.plos.org/10.1371/journal.pgen.1006199

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, et al. 2012. The genetic prehistory of southern Africa. Nat. Commun. [Internet] 3:1143. Available from: http://www.nature.com/doifinder/10.1038/ncomms2140

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data.Tang H, editor. PLoS Genet. [Internet] 8:e1002967. Available from: http://dx.plos.org/10.1371/journal.pgen.1002967

Plagnol V, Wall JD. 2006. Possible Ancestral Structure in Human Populations. PLoS Genet. [Internet] 2:e105. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16895447

Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH, Saykin AJ, Orro A, Lupoli S, Salvi E, et al. 2009. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer’s disease.Domschke K, editor. PLoS One [Internet] 4:e6501. Available from: http://dx.plos.org/10.1371/journal.pone.0006501

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. [Internet] 38:904–909. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16862161

Prokopovic V, Popovic M, Andjelkovic U, Marsavelski A, Raskovic B, Gavrovic-Jankulovic M, Polovic N. 2014. Isolation, biochemical characterization and anti-bacterial activity of BPIFA2 protein. Arch. Oral Biol. [Internet] 59:302–309. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0003996913003701

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature [Internet] 505:43–49. Available from: http://www.nature.com/articles/nature12886

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. [Internet] 81:559–575. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, et al. 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc. Natl. Acad. Sci. U. S. A. [Internet] 105:1596–1601. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0711467105

Racimo F. 2016. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. Genetics [Internet] 202:733–750. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26596347

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature [Internet] 505:87–91. Available from: http://www.nature.com/articles/nature12736

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.Susko E, editor. Syst. Biol. [Internet] 67:901–904. Available from: https://academic.oup.com/sysbio/article/67/5/901/4989127

Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, et al. 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am. J. Hum. Genet. [Internet] 94:496–510. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929714000676

Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D. 2017. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. [Internet] 45:D877–D887. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1012

Rausell A, Mohammadi P, McLaren PJ, Bartha I, Xenarios I, Fellay J, Telenti A. 2014. Analysis of Stop-Gain and Frameshift Variants in Human Innate Immunity Genes.Quintana-Murci L,

editor. PLoS Comput. Biol. [Internet] 10:e1003757. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25058640

Rosentul DC, Plantinga TS, Scott WK, Alexander BD, van de Geer NMD, Perfect JR, Kullberg BJ, Johnson MD, Netea MG. 2012. The impact of caspase-12 on susceptibility to candidemia. Eur. J. Clin. Microbiol. Infect. Dis. [Internet] 31:277–280. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21706251

Salas A, Richards M, De la Fe T, Lareu M-V, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A. 2002. The making of the African mtDNA landscape. Am. J. Hum. Genet. [Internet] 71:1082–1111. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929707604030

Saleh M, Mathison JC, Wolinski MK, Bensinger SJ, Fitzgerald P, Droin N, Ulevitch RJ, Green DR, Nicholson DW. 2006. Enhanced bacterial clearance and sepsis resistance in caspase-12-deficient mice. Nature [Internet] 440:1064–1068. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16625199

Saleh M, Vaillancourt JP, Graham RK, Huyck M, Srinivasula SM, Alnemri ES, Steinberg MH, Nolan V, Baldwin CT, Hotchkiss RS, et al. 2004. Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. Nature [Internet] 429:75–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15129283

Scheet P, Stephens M. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am. J. Hum. Genet. [Internet] 78:629–644. Available from: http://www.sciencedirect.com/science/article/pii/S000292970763701X?via%3Dihub

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. [Internet] 46:919–925. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24952747

Schlebusch CM, Gattepaille LM, Engström K, Vahter M, Jakobsson M, Broberg K. 2015. Human Adaptation to Arsenic-Rich Environments. Mol. Biol. Evol. [Internet] 32:1544–1555. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv046

Schlebusch CM, Jakobsson M. 2018. Tales of Human Migration, Admixture, and Selection in Africa.

Schlebusch CM, de Jongh M, Soodyall H. 2011. Different contributions of ancient mitochondrial and Y-chromosomal lineages in “Karretjie people” of the Great Karoo in South Africa. J. Hum. Genet. [Internet] 56:623–630. Available from: http://www.nature.com/articles/jhg201171

Schlebusch CM, Lombard M, Soodyall H. 2013. MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. BMC Evol. Biol. [Internet] 13:56. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-13-56

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Vicente M, Steyn M, Soodyall H, et al. 2017. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science (80-.). [Internet] 358:652–655. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28971970

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MGB, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science (80-.). [Internet] 338:374–379. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22997136

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature [Internet] 463:943–947. Available from: http://www.nature.com/doifinder/10.1038/nature08795

Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan J, Mägi R, Strawbridge RJ, Rehnberg E, Gustafsson S, et al. 2012. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat. Genet. [Internet] 44:991–1005. Available from: http://www.nature.com/articles/ng.2385

Shimomura Y, Aoki N, Ito M, Rogers MA, Langbein L, Schweizer J. 2003. Characterization of Human Keratin-Associated Protein 1 Family Members. J. Investig. Dermatology Symp. Proc. [Internet] 8:96–99. Available from: https://www.sciencedirect.com/science/article/pii/S0022202X15529485#!

Skeldon AM, Morizot A, Douglas T, Santoro N, Kursawe R, Kozlitina J, Caprio S, Mehal WZ, Saleh M. 2016. Caspase-12, but Not Caspase-11, Inhibits Obesity and Insulin Resistance. J. Immunol. [Internet] 196:437–447. Available from: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1501529

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, et al. 2017. Reconstructing Prehistoric African Population Structure. Cell [Internet] 171:59-71.e21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28938123

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. [Internet] 84:740–759. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929709001633

Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, Richards JB, Hammond N, Stolk L, Nica A, Inouye M, Hofman A, et al. 2009. Meta-Analysis of Genome-Wide Scans for Human Adult Stature Identifies Novel Loci and Associations with Measures of Skeletal Frame Size.Visscher PM, editor. PLoS Genet. [Internet] 5:e1000445. Available from: http://dx.plos.org/10.1371/journal.pgen.1000445

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, et al. 2016. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr. Protoc. Bioinforma. [Internet] 54:1–33. Available from: http://doi.wiley.com/10.1002/cpbi.5

Stöver BC, Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics [Internet] 11:7. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-7

Szpiech ZA, Hernandez RD. 2014. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol. Biol. Evol. [Internet] 31:2824–2827. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25015648

Szpiech ZA, Jakobsson M, Rosenberg NA. 2008. ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics [Internet] 24:2498–2504. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18779233

Takeuchi T, Adachi Y, Nagayama T. 2012. A WWOX-binding molecule, transmembrane protein 207, is related to the invasiveness of gastric signet-ring cell carcinoma. Carcinogenesis [Internet] 33:548–554. Available from: https://academic.oup.com/carcin/article-lookup/doi/10.1093/carcin/bgs001

Team RC. 2013. R: A language and environment for statistical computing.

Teleman AA, Chen Y-W, Cohen SM. 2005. Drosophila Melted modulates FOXO and TOR activity. Dev. Cell [Internet] 9:271–281. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1534580705002571

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. [Internet] 47:D506–D515. Available from: https://academic.oup.com/nar/article/47/D1/D506/5160987

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, et al. 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol. Biol. Evol. [Internet] 24:2180–2195. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msm155

Tobias PV. 1964. Bushman Hunter-Gatherers: A Study in Human Ecology. Ecol. Stud. South. Africa:69–86.

Traherne JA. 2008. Human MHC architecture and evolution: implications for disease association studies. Int. J. Immunogenet. [Internet] 35:179–192. Available from: http://doi.wiley.com/10.1111/j.1744-313X.2008.00765.x

Triska P, Soares P, Patin E, Fernandes V, Cerny V, Pereira L. 2015. Extensive Admixture and Selective Pressure Across the Sahel Belt. Genome Biol. Evol. [Internet] 7:3484–3495. Available from: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evv236

Troxler RF, Offner GD, Xu T, Vanderspek JC, Oppenheim FG. 1990. Structural Relationship Between Human Salivary Histatins. J. Dent. Res. [Internet] 69:2–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2303595

Turell MJ, Knudson GB. 1987. Mechanical transmission of Bacillus anthracis by stable flies (Stomoxys calcitrans) and mosquitoes (Aedes aegypti and Aedes taeniorhynchus). Infect. Immun. [Internet] 55:1859–1861. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3112013

Turner M. 1980. Anthrax in humans in Zimbabwe. Cent. Afr. J. Med. [Internet] 26:160–161. Available from: https://journals.co.za/content/CAJM/26/7/AJA00089176_1157

Turner WC, Imologhome P, Havarua Z, Kaaya GP, Mfune JKE, Mpofu IDT, Getz WM. 2013. Soil ingestion, nutrition and the seasonality of anthrax in herbivores of Etosha National Park. Ecosphere [Internet] 4:art13. Available from: http://doi.wiley.com/10.1890/ES12-00245.1

Veeramah KR, Thomas MG, Weale ME, Zeitlyn D, Tarekegn A, Bekele E, Mendell NR, Shephard EA, Bradman N, Phillips IR. 2008. The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa. Pharmacogenet. Genomics [Internet] 18:877–886. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18794725

Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. 2013. HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. Hum. Mutat. [Internet] 34:1189–1194. Available from: http://doi.wiley.com/10.1002/humu.22356

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol. Biol. Evol. [Internet] 26:1823–1827. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp096

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution (N. Y). 38:1358–1370.

Wilson MSC, Bulley SJ, Pisani F, Irvine RF, Saiardi A. 2015. A novel method for the purification of inositol phosphates from biological samples reveals that no phytate is present in human plasma or urine. Open Biol. [Internet] 5:150014. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25808508

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C. 2009. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. Am. J. Hum. Genet. [Internet] 84:224–234. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929709000159